Isolation and characterization of bacteriophages with therapeutic potential

Villarroel, Julia

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Ph.D. Thesis
Doctor of Philosophy in Bioinformatics

DTU Bioinformatics
Department of Bio and Health Informatics

Isolation and characterization of bacteriophages with therapeutic potential

Julia Villarroel

Kongens Lyngby 2018

DTU Bioinformatics
Department of Bio and Health Informatics
Technical University of Denmark

Kemitorvet, Building 208
2800 Kongens Lyngby, Denmark

www.bioinformatics.dtu.dk

There is a vitality, a life force, an energy, a quickening,
that is translated through you into action,
and because there is only one of you in all time,
this expression is unique.
And if you block it, it will never exist
through any other medium
and will be lost.

Martha Graham

Preface

This PhD thesis was prepared at the Technical University of Denmark, Department of Bio and Health Informatics (DTU Bioinformatics) in fulfilment of the requirements for acquiring the PhD degree. The research described was conducted under the supervision of Professor Morten Nielsen, Department of Bio and Health Informatics, Technical University of Denmark, and Ph.D. Mette Voldby Larsen, CEO of GoSeqIt and the co-supervision of Associate Professor Mogens Kilstrup, Department of Biotechnology and Biomedicine, Technical University of Denmark.

The first project HostPhinder stay in Argentina October and December 2014 were the main focus was the optimization of the HostPhinder tool, presented in this thesis in chapter 2.

The work presented in this thesis was carried out between december 2013 and october 2017 and was interrupted for ten and a half months of maternity leave.

The PhD was funded by DTU.

Kongens Lyngby, January 5, 2018

Julia Villarroel

Abstract

The concerning spread of antibiotic resistant bacteria has directed the spotlight upon bacteriophages, in short phages, as potential candidates for therapeutic purposes. Far for being a novelty, phage therapy has been widely used in the 20s and 30s in western countries until the discovery of antibiotics, which, coupled with a lack of knowledge of phage biology at that time, let to the replacement of phage therapy by antibiotics. On the other side of the planet, the Georgian Eliava Institute has been using phages for treating bacterial diseases since short after phage discovery a century ago. Georgian pharmacies commonly sell phage cocktails from the Institute without the need of a doctor's prescription. A thorough characterisation of the cocktail is though required for it to be accepted as pharmaceutical in the European Union. The potential to investigate the genetic material of microbial communities directly from the environment through metagenomics, allows for genomic characterisation of these cocktail. Furthermore, metagenomics analyses may lead to the discovery of novel phages with therapeutic potential, opening up a promising new horizon for phage therapy.

This thesis is divided into five parts, each assigned a chapter. Chapter 1 provides the reader with an introduction to phage biology, history and metagenomics. Here, the main bioinformatics methods used throughout the studies of the following chapters are also presented and briefly described. Chapter 2 presents the paper "HostPhinder: A Phage Host Prediction Tool" published in May 2016. The tool predicts the bacterial host of a given phage based on co-occurrent k-mers between a query sequence and reference phage genomes with known host. HostPhinder's accuracy in predicting the host species and genus of an evaluation set was higher than 74% and 81%, respectively. The tool can be applied to identify the host of phage sequences found for instance in metagenomes allowing for a first step characterisation. Chapter 3 presents the paper "Metagenomic analysis of therapeutic PYO phage cocktails from 1997 to 2014" submitted in October 2017 and currently under peer-revision. In this study, the compositions of 3 batches of a Georgian cocktail from 1997 to 2014 was compared by means of Next Generation Sequencing (NGS) and metagenomic analysis. Thirty and 29 phage draft genomes were found in the cocktails from 1997 and 2014, respectively. One of them was present in both sample and did not resemble any known phage genomes, strongly suggesting its novelty. Phage representatives of all bacterial targets supposedly targeted by the cocktail's were found, as predicted using HostPhinder. A comparison between cocktails from 1997, 2000, and 2014 showed a closer composition

between the first two cocktails. Chapter 4 presents the characterisation of historical S. aureus phages, once used for phage typing. Finally, the conclusive Chapter 5, recapitulates the main findings of this thesis and frame them into the perspective of potential future investigations.

Dansk Resumé

Den stigende bekymring for antibiotika i forbindelse med resistente bakterier har rettet blikket mod bakteriofager, eller blot fager, som en mulig kadidat for terapeutisk behandling. Langt fra at være en nyhed, fagterapi har været alment benyttet i 20'ern eog 30'erne i de vestlige lande indtil opdagelsen af antibiotika som kombineret med en manglede kedskab til fagbiologi på det tidspunkt, førte til antibiotika erstattede fagterapi. På den anden side af planeten havde Georgian Eliava Institute brugt fagterapi til at bekæmpe bakterielle sygdomme siden kort efter opdagelsen af fager for et århundrede siden. Georgiske apoteker sælger ofte fagcocktails fra instituttet uden en lægeanvisning. En omfattende karakterisering af en sådan cocktail er dog krævet for at blive accepteret som farmaceutisk i Den Europæiske Union. Potentiallet for direkte at kunne undersøge det genetisk materiale i det mikrobielle samfund fra et metagenomics miljø åbner op for genetisk klassifikation af disse cocktails. Yderligere vil metagenomics analyse måske føre til opdagelsen af nye fager med terapeutisk potentielle, som vil kunne åbne for en ny lovende horisont for fagterapi.

Afhandlingen er delt op i fem dele hver tildelt et kapitel. Kapitel 1 giver læseren en introduktion til fagbiologi, historie og metagenomics. Hoved bioinformatik metoderne, som er brugt igennem studierne, er her præsenteret og kort beskrevet. Kapitel 2 præsenterer artiklen "HostPhinder: A Phage Host Prediction Tool" udgivet i maj 2016. Værktøjet prædiktere den bakterielle host for et givet fag baseret på co-occurrent k-mers imellem en forespørgselssekvens og reference faggenomer med kendt host. HostPhinder's nøjagtighed i prædiktering af host art og slægt af et evalueringssæt var højere end 74% og 81%, respektivt. Værtøjet kan blive anvendt til at identificere en host til en fagsekvens, som f.eks. fundet i metagenomes, som vil kunne bruges til første trin i en klassifikation. Kapitel 3 præsenterer artiklen "Metagenomic analysis of therapeutic PYO phage cocktails from 1997 to 2014" indsendt i oktober 2017 og under revision i skrivende stund. I dette studie var sammensætningen af 3 parti af en georgisk cocktail fra 1997 til 2014 sammenlignet med metoder af Next Generation Sequencing (NGS) og metagenomic analyse. Tredive og 29 fag kladde genomer var fundet i disse cocktails fra 1997 og 2014 respektivt. En af dem var tilstede i begge prøver og lignede ikke nogen kendt faggenomer, hvilket kraftigt tyder på at den er ny. Fagrepræsentanter for alle bakterielle mål angiveligt målrettet af cocktailen blev fundet, som forudsagt ved hjælp af HostPhinder. En sammenligning mellem cocktails fra 1997, 2000 og 2014 viste en tæt sammensætning mellem de to første cocktail. Kapitel 4 præsenterer karakteriseringen af de historiske S. aureus

fager som engang blev brugt til fagtyping. Endeligt det konkluderende kapitel 5 rekapitulerer de store fund af denne afhandling putter dem i perspektiv til en potentiel fremtidig undersøgelse.

Acknowledgements

The accomplishment of these PhD studies would not have been possible without my supervisors Mette Voldby Larsen and Morten Nielsen. Thank you for giving me the opportunity to conduct this PhD and for the invaluable guidance and coaching during these years. You taught me not only how to be goal-oriented and practical in order to achieve results, but also how to handle harsh criticism from reviewers with objectiveness, diplomatically, and in a productive way. Thanks also for your constant feedback, which was crucial for the accomplishment of this dissertation on time.

The so dreaded work in the laboratory turned out to be bearable, almost pleasurable, thanks to Mogens Kilstrup's supervision. Thank you Mogens for making sure I had the basic equipment to conduct the experiments and for sharing enlightening biological knowledge.

A big thanks to the system administrators at DTU Bioinformatics: Peter Wad Sackett, Ali Syed, Rafał Wolanin, and John Damm Sørensen for irreplaceable technical support.

The paperwork can become a nightmare for researchers and can steal lots of time. Thanks to the administration at DTU Bioinformatics, especially Dorthe, Lone, and Marlene, for promptly dealing with bureaucracy and making sure paperwork worked with ease.

Working days at DTU Bioinformatics would not have been the same without brilliant, (sometimes a bit noisy ;)), office mates. Thank you, Marie, Kosai, Franzi, Jakob, Katrine, Henrike and Jose, for distracting me in times when I started taking things too seriously. Your companionship and willingness to reach out to help others in their achievements has been exemplary to me.

During my stay in Argentina, I had the pleasure of sharing the office (and drinking *mate*) with exceptionally welcoming and attentive researchers. Thanks Emilio, María Paula, Santiago, Raúl, and Massimo for making each single day at the Instituto de Investigaciones Biotecnológicas of San Martín memorable and for the great *asados* events we enjoyed outdoors. Your kindness, coupled with the day-by-day supervision by Morten, which I enjoyed so much, made these three months intensively enjoyable and productive.

This long and short journey would not have been possible without my family. Thanks to my parents, Edith and Guillermo, who have always believed in me and supported me in my aspirations both practically and morally. Thanks to my danish family, Grethe and Ivan for welcoming me in their hearts. Thank you Martin, for

your love, care and not the least for your help in organising the layout of this thesis. Thank you Freya, our daughter, for being such a joyful little girl. The feeling of urge to allow you to grow strong, confident and happy dissipates any worry that may occur during the day.

Contents

CHAPTER 1

Introduction

1.1 Background

1.1.1 Phage biology

Bacteriophages, phages, are viruses that predate bacteria and depend on bacterial metabolism and replication machinery to produce and transmit their progeny. The genetic material of phages consists of single-stranded or double-stranded RNA or DNA and ranges from the 3.5 kb of the ssRNA phage MS2 to the 500 kb of the dsDNA Bacillus phage G. The discovery of the electron microscope in the 1940s enabled the identification of different phage morphologies; phages can be tailed, polyhedral, filamentous or pleomorphic, and some have lipid or lipoprotein envelopes [19, 4]. Based on the genome and the morphology, the International Committee on Taxonomy of Viruses (ICTV) has since 1966 made efforts to taxonomically classify phages [3, 63]. The majority (96%) of the visualised phages are tailed and belong to the order *Caudovirales* (from latin cauda, tail), which bear double stranded DNA (dsDNA) as genetic material. The families in this order are *Myoviridae* characterized by a straight contractile tail, *Podoviridae* having a short tail and *Siphoviridae* with a non-contractile and flexible tail (Figure 1) [4].

Phages differentiate according to different reproductive cycles. Whilst obligately lytic phages lyse the bacterial cell upon infection (lytic cycle), temperate phages introduce their genetic material into the host as a prophage and either replicate in concert with the host DNA in a free plasmid like state or integrate into the bacterial chromosome (lysogenic cycle). Stress signals usually trigger the expression of lytic genes in temperate phages that eventually exit the bacterial cell by lysing it. The lysogenic cycle establishes a symbiotic relation that in some cases results in mutualism, where the prophage, the integrated phage, provide fitness and evolutionary advantages to the lysogenic bacterium. A third kind of phage cycle is undertaken by filamentous phages, which continuously secrete progeny in a chronic infection that slows the growth of the host cell (pseudo-lysogenic cycle).

The structure of typical *Myoviridae* has been thoroughly studied and described [15] and consists of the head, an icosahedral capsid, a tail usually covered by a sheath, a base plate and the tail fibers. Phage tail fibers recognise and bind to specific receptors or polysaccharides on the bacterial cell membrane. The phenomenon named adsorption determines the high specificity of receptor recognition that link phages

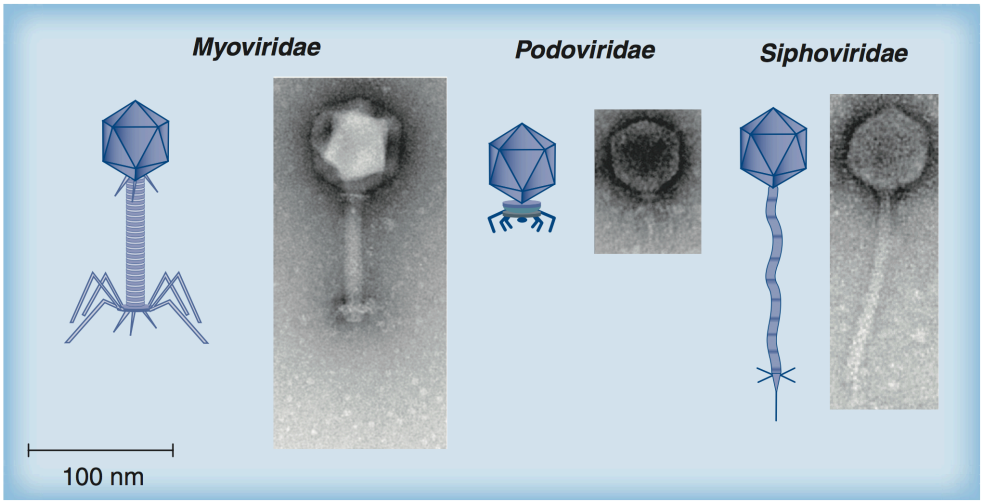


Figure 1.1: Families of the order *Caudovirales*. ©2011 D.R. Harper, J Anderson & M.C. Enright.

to specific host strains. Upon adsorption, the phage tail sheath undergoes a conformational change, which allows the quick passage of the genomic material from the capsid, through the hollow tail, and into the host cell [124, 62, 90, 91]. Phages and bacteria have coevolved in a fierce arms race that involved the development of strategic defence mechanism by bacteria and the phage counter attack with restored infectivity.

The mechanisms of resistance by bacteria target various stages of the phage infection, Figure 2. Bacteria can modify the receptor on the cell membrane responsible for phage recognition and adsorption. By the use of restriction enzymes, bacteria cut and destroy the phage genetic material that entered the cell. When infected, bacteria may activate the cascade that brings to bacterial apoptosis; this altruistic mechanism takes the name of abortive infection (Abi) [105, 25]. The use of restriction enzymes and the Abi are part of the bacterial innate immunity, i.e. rapidly activated and unspecific to the invading agent. The clustered regularly interspaced short palindromic repeats (CRISPR)-Cas mechanism, on the other side, is specifically directed to previously encountered foreign genetic elements, accounting for the adaptive immunity of the cell. The CRISPR array, in fact, keeps track of previous infections endured by the cell. When viruses infect the cell for the first time, the specific Cas proteins cleave the invading genetic material and introduce the resulting snippets into a CRISPR array. These snippets are used to recognize a second invasion from the same phage and to promptly inactivate the invading DNA/RNA [96].

The ever evolving coexistence of phages and bacteria is responsible for bacteria

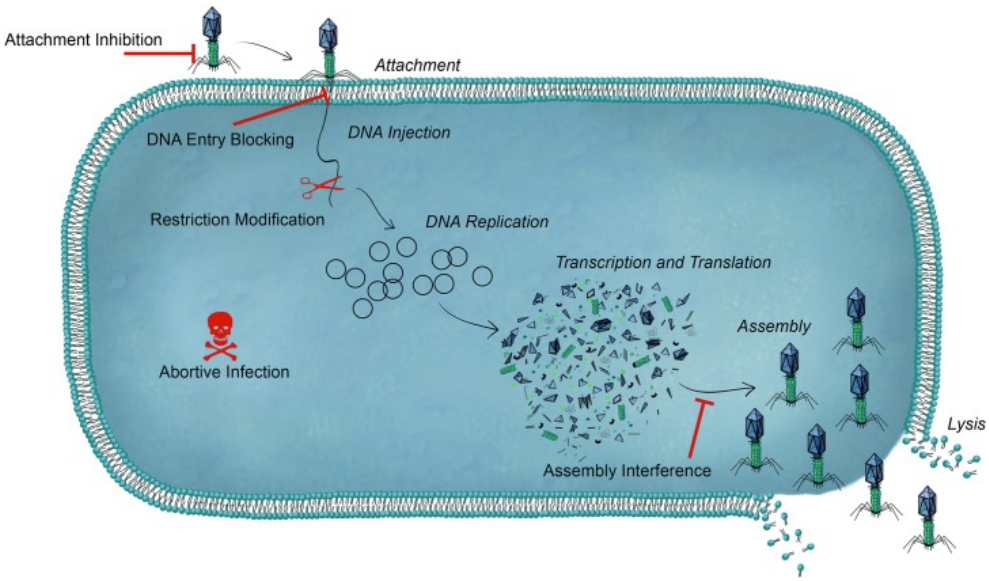


Figure 1.2: Bacterial Resistance mechanisms against phages. ©2015 Kimberley D. Seed.

diversification. Indeed, phages can give fitness advantages to bacteria by transferring new genes to the host, broadening, for instance, their host range. The phenomenon by which phages convey genes from a donor bacterial cell to a recipient one is called horizontal gene transfer (HGT) and is undertaken by temperate phages undergoing lysogenic or pseudo-lysogenic life cycles. HGT can promote the transfer of genes encoding the ability to degrade toxic compounds, antibiotic resistance or virulence factors [41], such as Vibrio cholera toxin conveyed by the filamentous phage CTXφ [117, 32] and E.coli Shiga toxin transferred by the lambdoid phage H-19B [78, 87].

1.1.2 Crucial role of phages in defining core biological principles

Phages drove the molecular biology revolution and were key in defining core biological principles such the establishment of the central dogma of molecular biology: information is sequentially passed from DNA to RNA to proteins [99]. Phages are ideal and tractable model systems to assess key biological questions. Lambda phage and T4 are the best characterized biological systems [22]. Thanks to the simplicity of their structure, a protein coat and internal DNA, the leader genetic contenders as carrier of the genetic information, Alfred Hershey and Martha Chase in the elegant experiment identified the DNA as genetic material [55].

Furthermore phages provided techniques (genome engineering, editing and se-

quencing and phage display) and reagents (restriction enzymes, phage T7 high fidelity DNA polymerase, integrases, recombinases, CRISPR) that underpin modern biology and synthetic biology. In the last decade, research of bacterial immunity against phage infection has yielded one of the most promising genome-editing molecular techniques, the CRISPR-Cas system [14]. Over 40 years ago the first ssRNA phage was sequenced in 1976 [40] and the next year the genome of ϕX174 was the first complete ssDNA phage to be sequenced with the Sanger method [102].

1.1.3 History of phages and phage therapy

The presence of an antimicrobial activity, which we now know was related to phages, was first reported in 1896 by Ernest H. Hankin in the Jumna and Ganges rivers. The British chemist described, not without surprise, the potential of these waters to stop the spreading of cholera in the villages close to the rivers [50]. It was not before 1915 that an article published in the renowned journal Lancet by British bacteriologist Frederick Twort described the existence of a transparent material, which produced glassy areas in Micrococci cultures. It was however not clear if it was a virus, an enzyme produced by the same bacteria, an amæba, or a minute bacterium. At that time the only evidence available of the presence of an "ultra-microscopic virus" was its pathogenicity towards higher organisms. Twort raised the hypothesis that non pathogenic viruses grew on bacteria, while pathogenic ones grew only on the infected animals [114]. Two years later, in 1917, the French-Canadian Félix d'Herelle labelled the cause of bactericidal activities against *Shigella dysenteriae* and other pathogenic bacteria and coined the name bacteriophage [35]. Highly motivated by a successful experiment against chicken typhus [110], he pursued the application of phages as therapeutic and prophylactic treatment in humans based on phage features of selectivity towards pathogenic bacteria and innocuousness towards human host cells [42]. D'Hérelle founded the Laboratoire du Bactériophage in France and started the production of the first commercial phage cocktails—Bacté- Coli-Phage, Bacté-Intesti-Phage, Bacté-Dysentérie-Phage, Bacté- Pyo-Phage and Bacté-Rhino-Phage, still produced at the George Eliava Institute of Bacteriophage, Microbiology and Virology, or Eliava Institute in the Georgia SSR [2].

Phages were soon welcomed in the clinical practises and conceived as magic bullet [108]: a drug that is able to go directly to its target, without damaging healthy human tissues. Phages were considered a panacea and the ultimate cure for most, if not all, diseases. During the 20s and 30s, companies started producing phage stocks, even before the biology of these agents was understood [2, 51]. In fact, it was only with the discovery of the electron microscope in 1939 by Helmut Ruska that the viral nature of phages was discovered and shown. Upon the realization that phage therapy was not a panacea, as it was hoped to be, the initial enthusiasm towards bacteriophages quickly faded in western countries. Phage therapy was displaced by the arrival of effective antibiotics, which had more controllable manufacturing, standard composition and determinable pharmacokinetic [51].

1.1.4 Antibiotic resistance and obstacles in transferring the "Stalinist cure" in western medicine

The spread of antibiotic resistance resulting from their abusive use is getting hardly controllable. The speed of production of novel antibiotics cannot keep up with the rate at which bacteria are developing and spreading resistance leaving patients without a cure [121]. This sets the search for an alternative as a matter of urgency. In the last few years interest in phage therapy has grown as testified by the number of papers published on the topic: 700 as of October 2017 (using "phage therapy" as a search term in the PubMed search engine). In Georgia, the Eliava Institute produces and commercializes phage cocktails, which are a mixture of phages targeting etiologic agents of specific diseases, such as intestinal disorders or skin purulent infections. In Poland, the Hirszfeld Institute of Immunology and Experimental Therapy provides phage treatment for patients that have been through numerous ineffective antibiotic treatments. Here phages are mainly used in synergy with antibiotics [64]. Yet the ''Stalinist cure'' as H. Brüssow refers to phage therapy [22], will not be taken seriously until controlled clinical trials of phage therapy are published in major medical journals. A detailed description of the composition of the drug may aid the approval of phage cocktails by the European Medical Agency and the US Food and Drug Administration (FDA). As biological entities, phages are programmed to mutate over time. Furthermore, phages multiply at the site of infection, which is a two sided coin: it has the advantage of effective treatment even at low doses, but on the other side it implies a complex pharmacokinetic that cannot be simply reconstructed. Controlled human clinical trials of phage therapy are required by the FDA, which also stresses the need of several phage therapy trials conducted against different infectious diseases [22]. One such trial is the Phagoburn trial, which launched on June 1st 2013 as a cooperative project between France, Belgium and Switzerland. This phase I/II clinical study aims at assessing the safety, effectiveness and pharmacodynamics of phage cocktails to treat burn wound infections.

1.1.5 Phages Metagenomics

Metagenomics refers to field of study which aims at characterizing the genetic material of environmental samples, metagenomes. The potential to investigate microbial communities directly taken from their environment, overcomes the problem of cultivation and the biases that may arise from it [113]. One of the goal of metagenomics is finding out what is present in the sample by recovering whole genome sequences. Another objective is to identify the functions that are coded in the metagenome by looking at which genes are present and how they influence metabolic pathways [31]. Metagenomics can also detect relations of dependence and coevolution between biological elements such as those existing between phages and their hosts [113].

Phages are the most abundant biological entities on earth, counting 10^30 particles in the biosphere. Viruses are estimated to be 2.5×10^8 per millilitre of seawater

(bergh1989high, mann2005third, wommack2000virioplankton) and 1.5×10^7 bacteriophages per gram of soil (ashelford2003elevated). It is perhaps unsurprising therefore that phages influence cycles of nutrients, organic carbon and other chemicals; an impact that have ultimately a global implication on the entire biosphere [118, 98]. Phages play also a central role within the human gut virome where several studies have revealed a correlation between specific phage abundances and patient health [74, 73, 75, 93, 95, 38, 79].

Full shotgun metagenomics is well suited to the study of phages, especially because it overcomes the limitation that more than 99% of the bacterial hosts are unculturable [30, 92]. Due to their small size of phages, though, viral DNA usually contribute to 2-5% of total DNA in metagenomic samples [94]. To improve the detection of viral DNA one may start by isolating the fraction of the sample that contains the free viral particles [23]. Since the first study on viral metagenome, virome in 2002 [21], the number of published viromics studies has risen to almost 600 as of end of 2016 [54]. Yet the number of phage genomes in public databases is dwarfed by their bacterial counterpart. In fact, despite their abundance and critical evolutionary role in ecology and human health, less than 1% of the extant viral diversity is represented in public databases [119, 76]. As a consequence, the amount of unclassified sequences in viral metagenomes can easily sum up to 50% of the annotations [97]. In general 60–99% of the sequences generated in viral metagenomic studies are not similar to known viruses, [76], which poses challenges for taxonomic labeling or host species prediction.

1.2 Methods

1.2.1 Methods to compare sequences

Sequence comparison between query and database is used in genomics to examine relatedness and determine the taxonomic group of the query sequence. Sequence alignment algorithms calculate sequence similarity between query sequence and database and compute the statistical significance. There are two kinds of sequence alignment: global or local. The global alignment forces the alignment to span the entire length of the query sequence and may include large stretches of low similarity. On the other hand, local alignment identifies regions of similarity within long sequences that may overall be widely divergent; the similarity score in this case is not influenced by unconserved regions. The most widely used local alignment algorithm is BLAST, Basic Local Alignment Search Tool.

Another local alignment algorithm is KmerFinder [53, 67], which is based on k-mers: sequence segments obtained by sliding a window of length k by one nucleotide at a time. The program is based on a database of k-mers, the keys, each linked to the sequence's ids of the genomes that contain the k-mer, the values. Upon searching for co-occurring k-mers between the query and the database the program assigns scores to database entries, calculates significance statistics, and outputs the most similar sequence to the query.

1.2.2 Data redundancy

Pairs of identical or highly similar sequences pervade public databases in a non homogeneous way: while some sequences are similar to many other sequences, other sequences are unique. This represents a concern, for instance in statistics when evaluating a model by cross validation (CV). In this case the sequences in the training and the evaluation set have to be different enough in order to not overestimate the performance of the method. In general both descriptive and predictive tasks benefit from training on a representative set of sequences devoid of duplicates.

The Hobohm I and II algorithms [56] were first developed in 1991 to reduce redundancy in the Protein Data Bank data sets, by automatically creating a list of representative protein structures with the highest diversity in sequence space, whilst uniquely representing each protein families. The simplicity and versatility of the algorithms make them applicable to any dataset in which similarity between data points can be computed. The Hobohm I algorithm starts by selecting the first sequence from a sorted list. Then each sequence in the sorted list is processed as follows: if similar to previously selected sequences is discarded, otherwise is selected. The process is reiterated until no sequence is left in the sorted list. The Hobohm II, upon linking each sequence to a list of neighbours (similar sequences), removes from the list the sequences with the largest number of neighbours. Then it updates the lists of neighbours of the remaining sequences and repeats. Whilst the first algorithm ("select until done") optimises based on a certain property by sorting the sequences according to that property; the second ("remove until done") maximises the number of selected sequences, and is unbiased by any prior sorting of the input data.

1.2.3 Model fitting, overfitting and model evaluation

Model fitting refers to the process of finding the model that best describes the observed data. The aim of creating a model on known data is to learn the underlying pattern, which can then be used to classify a new set of unknown data. Data are usually distributed around the function of the model and the distance between the model prediction for a given data point and the observed data point is called residual. Since residuals can be positive or negative, they are squared to get the respective magnitude. The model with the highest likelihood, the one that most likely generated the data, is the model with the smallest residual sum of squares.

To train and validate a model, two datasets are needed: the training set and the validating or test set. The training set is the set of data used to fit, train, the model. The error yielded by the model on the training set is referred to as training error. The test set is the set that is used to calculate the validation error, which is the error the model makes in predicting values for previously unseen data. The validation error best estimates the model predictive accuracy and hence has to be minimized.

In the process of model fitting shown in Figure 1.3, a model with very few independent parameters, degrees of freedom, may be too simple to comprehensively describe the data. This is referred to as underfitting. As the complexity of the model grows

and gains degrees of freedom, the training error steadily decreases. The validation error, on the other hand, will initially decrease, then reach a minimum and eventually increase again. The increase in the validation error is usually a consequence of overfitting. Overfitting occurs when the model learns the noise of the data and not the underlying trend of the observations. The model therefore, fits perfectly the training data but extrapolates poorly to unseen data.

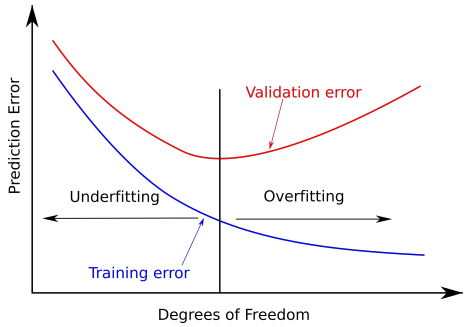


Figure 1.3: Prediction error versus model complexity in terms of degree of freedom. The optimal level of complexity of the model is at the global minimum of the validation error. Modified from ©User:Gringer / Wikimedia Commons / CC-BY-SA-3.0.

The data for the training and the test sets have to be drawn from the same population in order to get a reliable evaluation of the predictive performance of the model. As an example, consider a clinical setting where the modulatory effect of phages on the gut microbiota is to be predicted for patients of equal gender ratio and representing a wide age range, the test set. If the model is trained on a very specific group, say young men, the model will lack representative observations that are essential for estimating the effect of similar data in the test set; the predicting power of the model will therefore be very low. Similarly, training and test sets should be different enough in order to avoid falsely high performance. The problem linked to having identical or nearly identical observation in the training and test sets is that some of the data points in the training will have the same values of predictors as the observations in the test set leading to a falsely low validation error and poor external validity. This problem can be overcomed by homology reducing observations prior to data partitioning. Hobohm algorithms, previously described, can be used for this purpose.

1.2.4 Comparison of models

Predictive accuracy is the basis for comparing models. Models with higher predictive accuracy reflect better the law that generated the data and make more useful predictions.

It allows to compute the validation error and compare the predicting power of different models in order to select the best performing model. Cross validation (CV) allows to compare models by partitioning the dataset into folds, for example 4 in a 4-fold CV. In turn, 3 partitions are used as training set to fit the model that is then tested on the 4th partition, the test set. Then the next combination of training partitions is used on a second test set, and so on until, in rotation, all partitions get to be the test set once. This setup allows the methods to be trained and tested on all the data available. Finally the accuracy of the model is given by the overall accuracy of all CV rotations, the CV-performance. The CV is run for each model to be compared, and the model with highest overall accuracy is selected. In cases where the test data is used in combination with the training data to construct the optimal model, the CV-performance of the best performing method is not an unbiased predictor of the model performance. This is for instance the situation if the test set was used for early stopping (a training procedure where the model parameter optimization is stopped when the test performance is optimal), if the test set was used to do feature selection, or if the test set in any other way was used to define model hyperparameter. In these situation, a nested-CV is needed to achieve an unbiased assessment of the predictive power of a given model. In a nested-CV, an evaluation set of data that the model will never see until the final evaluation should be separated from the training-test set. The data is partitioned in groups, for example 5. One of the partitions, the evaluation set, is set aside. With the remaining 4 partitions, a 4-fold CV is launched as described above. Once the best model is selected through CV, its performance is eventually estimated on the evaluation set by training the model on all 4 partitions used for the CV. So the process can be divided in 2 steps: 1) model selection through CV and 2) best model accuracy assessment by training the model on the data used for model selection (training-test data) and predicting the evaluation set.

1.2.5 Next Generation Sequencing

Next Generation Sequencing (NGS) technologies refer to the shotgun sequencing methods, such as 454, Illumina, SOLiD and Ion Torrent, that have replaced the first-generation, single-molecule sequencing technologies such as Sanger [101]. Shotgun technologies are able to cheaply produce an enormous volume of data and allow to sequence entire genomes with a fairly high accuracy. The Illumina technology is by far the most popular method, thanks also to the relatively low cost per base pair, and it accounts for 90% of the world's sequencing data, as of September 2017 [29].

The Illumina technology involves a first step of template preparation: the DNA is fragmented by random shearing into segments < 1 kb long forming a sequencing library; the extremities of the newly formed segments are attached to known sequences, adaptors, which are reverse complement of universal primers. The resulting recombinant molecules, templates, are then immobilized on a solid surface. To this follows an amplification step by bridge-PCR and massive parallel sequencing through cyclic reversible termination (CRT). The CTR involves the iteration of the following steps

1) incorporation of a fluorescent nucleotide terminator, 2) imaging to determine the incorporated base, 3) washing of unbound nucleotides, 4) cleavage of the terminating group and fluorescent dye and restarting the cycle. This is parallelly performed for each clone of each template on the solid surface [72]. The resulting reads are usually 250 bp long.

Before proceeding to downstream sequence analyses, raw reads need to be checked for quality. Low-quality reads can in fact compromise the assembly of the reads into draft genomes and the conclusions derived from it. Fastqc [9] is a program that performs a number of different analyses on raw reads and creates a quality report that can be visualized by a browser. The report has graphical modules, each illustrating the results from each quality analysis. The program gives a quick overview of whether the reads have quality problems that need to be addressed before performing any downstream analysis. Beside a quick overview of the quality of the reads, fastqc can reveal read issues such as the presence of duplicates or artefacts derived from sequencing the adapters, used in Illumina to attach the template to the sequencing surface. Prinseq-lite [103] is a command line program that allows to filter reads according to quality scores, sequence duplicates, reads length and percentage of ambiguous bases. Furthermore, using this program, reads can be trimmed at both ends and remove poly-A/T.

1.2.6 Assembly

Assembly is the process of merging reads together to reconstruct original sequences. This can be performed in a supervised way, by aligning the reads to a reference genome or in an unsupervised manner, denoted *de novo* assembly. The most popular *de novo* assemblers are based on the De Bruijn graph [28] to merge overlapping reads into contigs. Reads are first split into all possible k-mers. In the de Bruijn graph, each k-mer is represented by an edge connecting the k-mer prefix and suffix, each represented by a node, Figure 1.4. The following k-mer, successive edge, is shifted from the previous one by one position and will have as prefix the suffix of the previous k-mer and so on. As an example the first k-mer in Figure 1.4, ACCG has prefix ACC and suffix CCG and is followed by k-mer CCGT. Sequencing errors result in bulges in the de Bruijn graph, which are made of two path of similar length: the correct one, having higher coverage and the wrong one with a lower coverage, depicted in red in Figure 1.4 [126]. Assemblers adopt different algorithms for deconvoluting complex graphs and removing low-coverage alternative paths in a process called bulge removal [24, 82, 85, 126, 12]

A good single genome assembler produces the few, large contiguous contigs, ultimately one single contig that represents the entire genome. The assessment of the quality of the assembly, can be reference-based by assessing how well the assembly reproduce the reference sequence. However in most genomic sequencing experiment a reference genome is not available and the validation of the assembly must rely on *de novo* methods. The most commonly used *de novo* measure of assembly accuracy

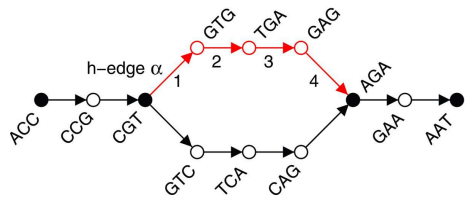


Figure 1.4: Depiction of the de Bruijn graph of fragments ACCGT**C**AGAAT and ACCGT**G**AGAAT using a k = 4. [12] ©2012, Mary Ann Liebert, Inc. An incorrect read resulting from a sequencing error, produces a bulge in the Bruijn graph, here displayed by the red edges 1, 2, 3 and 4. In a metagenome a bulge may be the result of nucleotide variances or indels of a rare strain compared to the most common strain..

is the contiguity measure N50: the length of the smallest contig in the set of largest contigs that together contain at least 50% of the full assembly [66].

1.3 Metagenomics pipeline

The analysis of metagenomic NGS data is in many ways similar to the analysis of single genome NGS data. One critical difference however is linked to the presence in metagenomes of genetic material from many biological entities. This poses challenges in defining for example what is present in the samples. Only a small fraction of sequences are in fact known and published, which limits the taxonomic labeling of the sequences in the sample by alignment to public data. As of the reconstruction of draft genomes, besides assembling reads into contigs, a further challenge is the need to group the contigs that originate from the same original genome.

Several tools have however been developed aiming at resolving these challenges

1.3.1 Reads mapping

Reads mapping, the alignment of reads to reference genomes, allows to gain an overview of what is present in metagenomic samples.

In Kraken [120] mapping is based on k-mers exact matches between a read and a database. The database is made of k-mers, each linked to the taxonomic tree of organisms that contain that k-mer. When aligning a read, Kraken assigns each k-mer in the read to the lowest common ancestor (LCA) of all the organisms in the database, whose genome contain the k-mer. In the taxonomy tree each node gets as weight the number of k-mers present in the genome that corresponds to the node's taxon. These weights add up to the score of each Root-To-Leaf (RTL) path. The taxon leaf of the RTL path with the highest score gives the taxonomic label to the read.

Another tool that performs read mapping is MGmapper [84]. Here, the taxonomic annotation of the database is given by a hash, where each reference genome identifier, the key, is linked to the full taxonomy path, the value. MGmapper assigns a taxonomic group to fastq sequence reads by aligning them to the reference sequences in the database using BWA [70]. Therefore, differently from Kraken, which looks for exact alignment of k-mers, MGmapper allows for indels and mismatches.

1.3.2 Assembly of metagenomes

High throughput shotgun sequencing is particularly suitable to study metagenomes. The assembly of metagenomes poses several issues that have to be addressed by assembly algorithms: metagenomes often contain a mixture of highly related bacterial strains that share most of the genomic sequence, though they present differences as a results of mutations, mobile elements insertion, horizontal gene transfer, and genome rearrangements. In de Bruijn graphs of metagenomes not only sequencing errors result in bulges, but also mutations or indels in rare strains compared to the most abundant strains. In this case the alternative, low-coverage path corresponds to the rare strain and the high-coverage path coincides with the abundant strain. This can only hardly be assessed by *de novo* algorithms, which yield fragmented reconstruction of metagenomic genomes from short-reads libraries. Metagenome assemblers usually discard information about rare species or variants to improve assembly contiguity [83, 69]. As an exception metaSPAdes keeps rare species and strains information, by just disconnecting low-coverage paths from the main path without removing them [80].

As of quality assessment of metagenome assemblies, the contiguity measure N50, presented above is not well suited. In metagenomic samples, multiple biological entities are present at different relative abundances and a measure of contiguity would be too simplistic. Other *de novo* measures such as ALE, assembly likelihood evaluation [26] and LAP, **l**og **a**verage **p**robability [45] estimate the likelihood of an assembly given the reads, which reflects how well the reads agree with the assembly. By simulating the sequencing process and taking into account the sequencing errors deriving from it, these measures estimate the probability of observing the reads from the assembled sequence. These are not absolute measures, but can be used to compare different assemblies generated from the same reads. These measures are specially suited to assess the correctness of metagenomic assemblies by scoring them according to how correctly they reproduce the sequences and the relative abundances of the individual biological entities [26].

1.3.3 Binning

After assembly, the next step in deconvoluting metagenomes is clustering related contigs in a process called metagenomic binning. Measures such as GC-content and the top Blast hits against publicly available sequences have previously been used to achieve this task [44, 127]. Both have intrinsic problems: the GC-content can vary

considerably within one genome, and no close relatives may be present in the database. The latter issue, linked to the Blast-approach limitation, is especially troublesome in analysing metagenomes, where the discovery of new species and functions is the main objective. An unsupervised partitioning of contigs based on shared genomic signatures therefore seems like a better suited solution for metagenomic data.

The genomic signatures of oligonucleotides of different lengths have been compared by means of chaos game representation [34], naïve bayesian classifier [100], and unsupervised neural networks [1], which found tetranucleotide frequencies (TNFs) as good signal of sequence relatedness [111]. The biological significance of TNFs has been linked to the presence of restriction enzymes specific towards palindromic 4-base segments. Specific tetranucleotide were found to be under-represented in bacteria that encode 4-base restriction enzymes that cleaved at these tetranucleotides. It has also been suggested that species-specific tetranucleotide preferences may be linked to the characteristic DNA synthesis and repair enzymes and the binding sequence preferences of transcription factors [1]. Examples of binning tools that utilizes tetranucleotide frequencies are TETRA [112], ESOM [36], Metawatt [109], GroopM [57] and MaxBin [122].

When multiple samples are available, co-abundance of contigs across samples of the same type has also been used to segregate contigs into those that derive from distinct biological entities. Examples of binning algorithms using inter-samples co-abundance are Canopy [79], CONCOCT [6], and MetaBAT [61]. Specifically, MetaBAT, (Metagenome Binning with Abundance and Tetra-nucleotide frequencies) is an unsupervised binning algorithm that combine information from both TNF and contig abundance probabilities of contigs derived from shotgun metagenomic. The tetranucleotide frequency probability distance (TDP) and abundance distance probability (ADP) across multiple samples are calculated between each pair of contigs. The TDP and ADP combined for each pair provide a distance matrix for all pairs. The distance matrix is used to cluster the contigs using a modified k-medoid algorithm that does not need an input number of medoids, k; hence the number of bins is not fixed *a priori*, but rather new bins are added until no contigs are left. Thereafter only large bins are kept, while the remaining ones are disaggregated in single contigs. If more than 10 samples are available the free contigs are joined to existing bins based on abundance correlation across the samples.

1.3.4 Taxonomy annotation

One of the goals of metagenomic sequencing is finding out what is presence in the sample by identifying biological entities and classifying them taxonomically. Probably the simplest and most effective way to assign metagenomic contigs to taxonomic groups is to compare them to known sequences. Of course, if the sequence shares no similarities with what has been previously seen, it cannot be classified as other than novel. Blast can assign a taxonomic label to unknown sequences by finding the best alignment to a large database of classified genomic sequences. Blast can therefore be

easily adapted to taxonomic labelling even though it was not originally designed for metagenomic analysis [20].

MetaPhinder [59], is a Blast-based method to identify sequences in metagenomes, which are of phage origin. It is based on a curated Blast database of publicly available phage genome sequences. To account for the modularity and rearrangements of phage genomes, instead of using the similarity value to the best hit, MetaPhinder computes the overall similarity to all significant blastn hits (E-value ≤ 0.05), Figure 1.5.

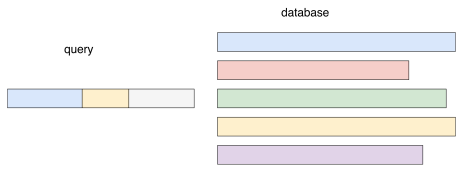


Figure 1.5: MetaPhinder calculates the **a**verage **n**ucleotide **i**dentity, ANI, from the alignment of the query to all significant blastn hits and from the query coverage by all hits. Reprinted with the kind permission of Vanessa I. Jurtz..

The similarity measure used by the method is the **a**verage **n**ucleotide **i**dentity (ANI) calculated as follows:

$$\mathrm{ANI} = \frac{\sum_{i=1}^{n} \mathrm{id}_i \mathrm{al}_i}{\sum_{i=1}^{n} \mathrm{al}_i} m_{\mathrm{cov}} \tag{1.1}$$

where n is the number of Blastn hits between the query sequence and all sequences in the database with an e-value of 0.05 or smaller, id is the Blastn % identity value between the query and a given database hit, al is the corresponding blastn alignment length, and m_{cov} is the coverage of the query sequence over all hits. The simplicity of the methods allows for flexibility in application. One application is assigning a taxonomic classification to an unknown phage sequence by looking for the closest reference genome in the databases. In this application a Blast database is created from the unknown phage sequence, which is then searched with each phage genome in the databases. The search will result in a ANI value for each reference phage genome that reflects how well the reference genome is covered by the unknown sequence. The reference genome with the highest ANI gives the taxonomic classification to the unknown phage sequence.

CHAPTER 2

Prediction of phage bacterial hosts

With the advent of the genomic era, deciphering new phage genomes has become possible. The amount of data that can be created in a small amount of time poses now the problem of deciphering those data and extract useful information from it. In the case of phage discovery, lab methodologies allow now to extract and purify phage particles [23] and enrich for viral DNA. Once the phage DNA has been extracted and sequenced, one need to discriminate between phage or non phage reads and contigs; a task that can be performed by tools such as MetaPhinder [59]. Finally, contigs deriving from the same original sequence need to be binned into phage contigs into phage genome drafts or phage families. The first step that will help us characterizing the phage is finding out the host they are capable of infecting.

We developed HostPhinder to predict the bacterial host of phages. HostPhincer is trained on a database of downloaded publicly available phage genomes for which the host is known and reported on the GenBank entry. Upon giving a phage genome of interest as query, HostPhinder will look for the closest reference in the database and output the corresponding host.

The task of predicting the procaryotic host of a phage can be thought of as a classification problem, where the phage genome has to be classified into a class, the host bacterial species. To prevent overfitting during parameter tuning, a nested-CV approach was used. Here, upon homology reduction using Hobohm I [56], phage genome sequences, were split into 5 partition, one of which was left apart until final evaluation. Parameter selection and tuning was performed using the remaining 4 folds as training-test dataset.

The main limitation of this approach is its dependency on publicly available sequence representatives. It has been reported, as of 2014, that only 8 of the 29 bacterial phyla have phage representatives on public databases [17], therefore HostPhinder will be limited to this set of host until new phages will be added. HostPhinder could be applied in metagenome studies to identify the bacterial host of phage draft genomes.

viruses

MDPI

Article

HostPhinder: A Phage Host Prediction Tool

Julia Villarroel [1,*], Kortine Annina Kleinheinz [1], Vanessa Isabell Jurtz [1], Henrike Zschach [1], Ole Lund [1], Morten Nielsen [1,2] and Mette Voldby Larsen [1,*]

[1] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; kortinekleinheinz@gmx.de (K.A.K.); vanessa@cbs.dtu.dk (V.I.J.); henrike@cbs.dtu.dk (H.Z.); lund@cbs.dtu.dk (O.L.); mniel@cbs.dtu.dk (M.N.)

[2] Instituto de Investigaciones Biotecnológicas, Universidad de San Martín, CP(1650) San Martín, Prov. de Buenos Aires, Argentina

* Correspondence: juliavi@cbs.dtu.dk (J.V.); metteb@cbs.dtu.dk (M.V.L.); Tel.: +45-4525-2425 (M.V.L.); Fax: +45-4593-1585 (M.V.L.)

Academic Editor: Rob Lavigne
Received: 23 December 2015; Accepted: 19 April 2016; Published: 4 May 2016

Abstract: The current dramatic increase of antibiotic resistant bacteria has revitalised the interest in bacteriophages as alternative antibacterial treatment. Meanwhile, the development of bioinformatics methods for analysing genomic data places high-throughput approaches for phage characterization within reach. Here, we present HostPhinder, a tool aimed at predicting the bacterial host of phages by examining the phage genome sequence. Using a reference database of 2196 phages with known hosts, HostPhinder predicts the host species of a query phage as the host of the most genomically similar reference phages. As a measure of genomic similarity the number of co-occurring k-mers (DNA sequences of length k) is used. Using an independent evaluation set, HostPhinder was able to correctly predict host genus and species for 81% and 74% of the phages respectively, giving predictions for more phages than BLAST and significantly outperforming BLAST on phages for which both had predictions. HostPhinder predictions on phage draft genomes from the INTESTI phage cocktail corresponded well with the advertised targets of the cocktail. Our study indicates that for most phages genomic similarity correlates well with related bacterial hosts. HostPhinder is available as an interactive web service [1] and as a stand alone download from the Docker registry [2].

Keywords: "host specificity"; prediction; genome; k-mers

1. Introduction

In 2012, the World Health Organization (WHO) announced the beginning of the end of the antibiotic era, and the possible return to a time when even trivial bacterial infections could turn out to be fatal [3]. Since then, the problem of antimicrobial resistance has continued to grow and in the foreword to the WHO report "Antimicrobial resistance: global report on surveillance 2014" it is stated that "A post-antibiotic era-in which common infections and minor injuries can kill-far from being an apocalyptic fantasy, is instead a very real possibility for the 21st century" [4]. As emphasized by WHO there is an urgent need for treatment alternatives, one such being bacteriophages (phages). The idea of using phages for the treatment of bacterial infections dates back to 1919, when French-Canadian microbiologist Félix d'Herelle used them for treating a patient with severe bacillary dysentery [5]. For a number of historical reasons, phage therapy never became general practice in the West, although it has been used extensively in countries from the former Eastern bloc [6–9]. Several recent studies from the West have also demonstrated the effectiveness of phages as antibacterial treatment [10–13], and more countries are currently revisiting phage therapy [14,15]. Phages have furthermore been suggested for use in the agriculture and food industries [16,17]. Examples include their use for reducing *Campylobacter jejuni* colonisation of broiler chickens [18] and the growth of *E. coli* in milk [19].

For a phage to successfully infect a bacterial host, the phage must adsorb to the bacterial surface through recognition of specific host receptors, e.g., proteins, LPS, or cell wall polysaccharides. Phage adsorption to an appropriate surface receptor is, however, only the first step required for successful infection. Several host defence mechanisms must also be overcome: Restriction-Modification (RM) systems have been shown to be present in more than 90% of sequenced bacterial genomes [20]. These systems include restriction enzymes that degrade incoming phage DNA with appropriate target sequences. Some bacteria contain Clustered Regular Interspaced Short Palindromic Repeats (CRISPR) loci, which together with the CRISPR-associated (cas) genes encode an adaptive anti-phage immune system [21]. Phage abortive systems (Abi systems) allow infected bacteria to commit "altruistic suicide" thereby preventing the spread of the phage within the bacterial community [22]. Other factors such as successful gene transcription and translation based on amino acid or tRNA availability further limit the host range [23]. Bacteria and phages have from the outset of their coexistence been engaged in a vehement arms race leading to intricate coevolutionary processes, and for each of the defence mechanisms mentioned above, examples exist of phages that have evolved to circumvent them [24,25]. The arms race has contributed to bacterial as well as phage diversity [26] and entails that phage host determination is influenced by multiple genes and genome features distributed across the phage genome. Although examples exist of phages that have extended their host range based on only a few mutations [27], the extended host range is typically limited to different strains of the same species. Apart from polyvalent enterobacteria phages, which are able to infect members of phylogenetically linked genera within the *Enterobacteriaceae* family, e.g., *Escherichia*, *Shigella*, and *Klebsiella* [28,29], most phages have been found to be specific to a particular genus [30]. This has been indicated by studies examining proteins, not entire proteomes [31], as has the "Phage Proteomic Tree", which is based on completely sequenced phage genomes [32], and analysis of genome type for Mycobacteriophages and host preference [33].

In this study, we extend the observation that genetically similar phages often share the same bacterial host species and hypothesize that it should be possible to predict the host species of a phage by searching for the most genetically similar phages in a database of reference phages with known hosts. In the developed method, called HostPhinder, genetic similarity is defined as the number of co-occurring k-mers between the query phage and phages in the reference database. K-mers are stretches of DNA with a length of k, and their use as a measure of genetic relatedness dates back to Woese and Fox and their groundbreaking paper from 1977, which uncovered Archaea as a separate branch in the tree of life [34]. Woese and Fox limited their analysis to k-mers (they used the term oligonucleotides) in 16S (18S) ribosomal RNA, but since phages do not have 16S rRNA genes or any other genes which are common to all phages [32], and because high-throughput sequencing methods have made the entire genome of phages easily available, HostPhinder examines the complete genome. Further, for bacteria we have previously shown that the co-occurrence of k-mers across the entire genome performs superior to other whole-genome or single locus based approaches for inferring genetic relatedness [35]. The splitting of entire phage genomes into overlapping k-mers may furthermore be an advantage in relation to the highly mosaic phage genome structure [36,37].

We believe that a method enabling prediction of the bacterial hosts of phages will be useful for several reasons. Firstly, phages have for many years been used to treat bacterial infections in countries belonging to the former Eastern bloc. The Eliava Institute in Tbilisi, Georgia has in particular been dominant in this regard and produce cocktails containing a mixture of phages for a range of bacterial infections. One of the steps towards adopting phage therapy in the West, is likely to be a full characterization of the content of these cocktails, which due to the way they are manufactured is not known [38]. Further, the current approach to exploration of many ecological niches is done by untargeted sequencing of samples isolated directly from the environment, so called metagenomics. This enables identification of phage and bacterial sequences without knowledge of the link between them, and importantly also enables identification of bacteria, and hence phages, that cannot be cultured. HostPhinder could help establish the link between phages and bacteria, which might be an important

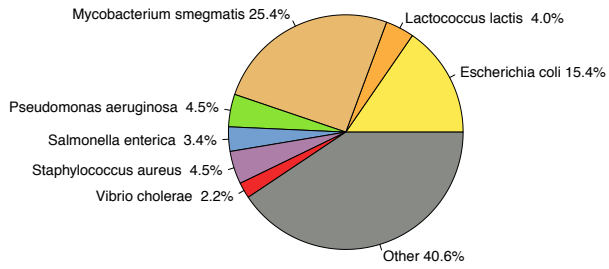
step towards understanding, e.g., the microbiome of the human gut, and possibly associations between the microbiome and clinical parameters of the human host [39].

2. Materials and Methods

2.1. Whole Genome Phage Sequences from Public Databases

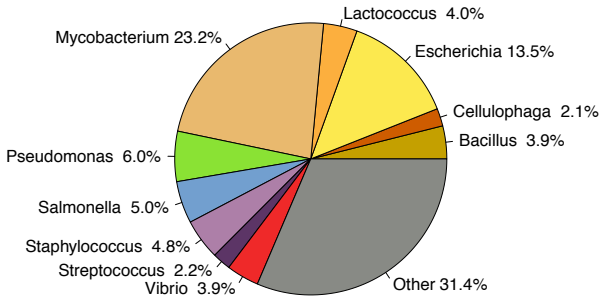
A set of public phage Whole Genome Sequences (WGS) was collected in August 2014: First, lists of phage WGS IDs were obtained from Phages.ids–VBI mirrors page [40], the NCBI viral Genome Resource [41], the EMBL EBI phage genomes list [42], and the phagesdb databases for Mycobacteriophages [43], Arthrobacter [44], Bacillus [45], and Streptomyces [46]. The resulting unique list of IDs was uploaded to the Batch Entrez service of NCBI to retrieve the corresponding WGS. Furthermore genome sequences were downloaded from the PhAnToMe genomes database and from NCBI searching for "(phage [Title]) AND complete genome".

Only entries indicating "complete genome" in the DEFINITION field of the GeneBank file and which host taxonomy was specified at least at the genus level were included. Entries annotated as "prophage" in the DEFINITION were removed. Hosts annotated as *Salmonella Typhimurium* were re-annotated as *Salmonella enterica* according to current nomenclature [47]. Finally, only the genus was taken into account for hosts with species specified as "sp." followed by an alphanumeric code; for example *Synechococcus* sp. *WH7803* was re-annotated as *Synechococcus*. 2196 phages had annotated host genus, here called $\text{phages}_{\text{genus}}$ dataset, and of these, 1871 had annotated species as well, $\text{phages}_{\text{species}}$. A total of 209 different host species and 129 different genera were represented among the phages (this data is available in HostPhinder's repository [48]). Figure 1 shows the distribution of hosts in the dataset.



(**a**) Distribution of host species.

Figure 1. *Cont.*



(**b**) Distribution of host genera.

Figure 1. Hosts represented in the database. Species (**a**) and genera (**b**) representations are displayed in the same genera-colour code.

2.2. Data Partitioning and Clustering

In this study, a 4-fold cross validation setup was used to assess the ability of the host prediction method to generalize to previously unseen data. Five data partitions were made, and one partition, $\text{phage}_{\text{eval}}$ was left aside during the entire process of parameter optimization. Once the parameters were optimized, the prediction accuracy was evaluated on this $\text{phage}_{\text{eval}}$ set, using the entire $\text{phage}_{\text{train,test}}$ set as reference database (Supplementary Materials Figure S1). In this setup, the performance of the evaluation set is hence completely unbiased towards the model parameter optimizations.

A reliable, *i.e.*, not overfitted, evaluation can only be made if phage genomes in the training-test and evaluation sets are not too similar to each other. Indeed, if a phage genome in the training set is almost identical to a genome in the evaluation set, it would be a simple task for HostPhinder to predict its host, leading to an overestimation of the method's ability to generalize to previously unseen data. To avoid such a bias we clustered the genomes according to 16-mer similarity by means of a Hobohm 1 approach [49]. The Hobohm approach consists in the formation of a final list of representative phage genomes, here called seeds. After the first sequence in a randomly sorted list enters the seed list and forms a seed, the following sequences are each checked for similarity (number of overlapping 16-mers) to each seed in the final list. Only if significantly different to the seed sequences, the new sequence will be included in the seed list. Otherwise, it will be linked to the most similar seed as member of the same cluster. The similarity between two genomes was measured in terms of frac_q (see Equation (4) in section "K-mer-based resemblance measures") using a threshold $\text{frac}_q > 0.7$. This threshold was chosen because the resulting clustering was most similar (93%) to the clustering obtained with a BLAST-Hobohm1 approach, where the similarity threshold was set to >90% genomewide ID (data not shown). The k-mer-Hobohm1 analysis resulted in 293 clusters with at least 2 sequences and 1121 singlets. The total number of seeds was hence 1414 containing 1 to 97 sequences. To separate the clustered phages in train-test and evaluation sets, the 1414 seeds were sorted by host alphabetical order, and secondly by size and alternately distributed between 5 partitions. This assured an equal host and genome size representation among partitions. Finally remaining members of each cluster were integrated into the partition of their respective seed. Sequences within the same cluster shared the host; therefore the unbiased host distribution was maintained also after integrating members of the clusters in each partition (see Supplementary Materials Figure S2). Subsets of each of these partitions were made, which comprised all phages that contained information about the species of the host, overall constituting the $\text{phages}_{\text{species}}$ dataset. The host and size distribution between partitions remained conserved (see Supplementary Materials Figures S2–S4). As stated above, one partition was next left aside for final evaluation, $\text{phages}_{\text{eval}}$, and the remaining 4 formed the train-test set,

$\text{phages}_{\text{train-test}}$. The final $\text{phages}_{\text{train-test,genus}}$ set contained 1818 phages (115 genera and 190 species), the $\text{phages}_{\text{eval,genus}}$ set contained 378 phages (72 genera, 96 species), while the $\text{phages}_{\text{train-test,species}}$ set consisted of 1546 phages and the $\text{phages}_{\text{eval,species}}$ set consisted of 325 phages (data available in HostPhinder's repository [48]).

2.3. K-mer-Based Resemblance Measures

Under the assumption that phages infecting the same bacterial host share genomic features, the host of a query phage should be predictable by searching for the most genomically similar phages in a reference database of phages with annotated hosts. The reference database was build from phage genome sequences and their reverse complements by splitting both into k-mers and sliding a window of length k along the sequences with step-size 1.

Query sequences were likewise split into k-mers, and for each reference sequence i having at least one k-mer in common with the query, a score, S_i, was defined as the number of identical unique k-mers between query and template. This score was subsequently used to determine the expectation value E_i:

$$E_i = N_{\text{Hits}} \frac{l_{u,i}}{L_{u,\text{tot}}} \tag{1}$$

where N_{Hits} is the sum of scores over all references, $l_{u,i}$, is the total number of unique k-mers found in the reference sequence i and in its reverse complement and $L_{u,\text{tot}}$ is the sum of unique k-mers over all references in the database. This expectation value was used to obtain a z-score:

$$z_i = \frac{S_i - E_i}{\sqrt{S_i + E_i + \eta}} \tag{2}$$

with $\eta = 0.001$ being a pseudocount used to avoid division by zero. Using SciPy, a two-sided p-value was generated from the z-score. All p-values were corrected using the Bonferroni method [50] by multiplying each p-value by the number of reference phages in the database:

$$p_{\text{corr}} = p_i * N_{\text{ref}} \tag{3}$$

where N_{ref} is the number of reference sequences in the database. HostPhinder outputs only significant hits, *i.e.*, $p_{\text{corr}} < 0.05$. Additionally, the values $\text{frac}_{q,i}$ and $\text{frac}_{d,i}$ were estimated. They represent the ratio of the score and the number of unique k-mers in query and reference sequences respectively:

$$\text{frac}_{q,i} = \frac{S_i}{q_{u,i} + \eta} \tag{4}$$

where $q_{u,i}$ is the number of unique query k-mers and $\eta = 0.001$ avoids division by zero. The value of $\text{frac}_{q,i}$, falling between 0 and 1, gives a direct indication of how much of the query sequence matched to the reference phage.

$$\text{frac}_{d,i} = \frac{S_i}{l_{u,i} + \eta} \tag{5}$$

where $l_{u,i}$ is the number of unique k-mers in the reference sequence and in its complement. Therefore, $\text{frac}_{d,i}$ falls between 0.5 and 1 if query and reference are identical, depending on the number of additional unique k-mers found in the reversed complement. The two measures are hence not directly comparable. Finally the coverage was determined as a measure of how much of the reference sequence is covered by the total number of k-mers in the query that match the reference:

$$\text{coverage}_i = \frac{2q_{\text{matched},i}}{l_{u,i} + \eta} \tag{6}$$

where $q_{\text{matched},i}$ is the total number of k-mers in the query that were matched to reference i, and l_i is the total number of k-mers in the reference. Both of these values include identical k-mers and do not only

count unique k-mers. The factor 2 is included to account for the additionally used reverse complement sequence of the reference to obtain l_i. The coverage can be larger than 1 if the query contains k-mers that could be matched multiple times.

2.4. Determining the Measure and Selection Criteria for Final Prediction

As described above, 5 measures were calculated for the similarity of a query phage to each of the phages in the reference database: score, z-score, frac_q, frac_d, and coverage. The optimal measure was determined in a simple 4 fold cross-validation setup. Here in turn, 3 of the 4 data sets were used as reference database for predicting the host for each query phage in the left out test set (see Supplementary Materials Figure S1, left). The host was inferred from the host of the reference phage with the highest value of similarity measure. This was repeated 4 times so that all 4 partitions were used as test set, and an overall performance for the given measure was calculated by concatenating the predictions of the 4 test sets. For each measure the average and interval of confidence was assessed through 100 bootstrap resamplings with replacement for each test set and calculating the overall accuracy. On a pairwise comparison based on 1000 bootstrap resamplings, coverage outperformed the other measures and was therefore chosen for further analysis. A number of different selection criteria can be used for the final prediction of the host of a query phage. We tested and compared the efficacy of 4 selection criteria that are each described in detail below.

2.4.1. Criterion 1: Host of Best-Matching Reference Phage

The host of the reference phage with the highest coverage value was selected as predicted host. This is the selection criterion used above to define the optimal similarity measure.

2.4.2. Criterion 2: Majority Host among Top-10 Reference Phages

As predicted host, the most abundant host among the hosts of the top 10 reference phages with the highest coverage values was selected. In case of a tie, the most abundant host with the highest coverage, was selected.

In cases where the coverage of non-top reference phages is far below the coverage of the top reference phage, it might not be advantageous to consider them in the selection criterion. To accommodate this, two additional criteria, criteria 3 and 4, were developed.

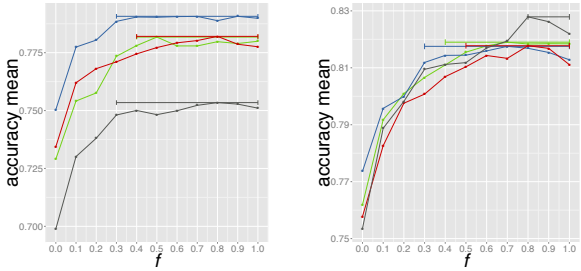
2.4.3. Criterion 3: Majority Host among Reference Phages above Coverage Threshold

As predicted host, the most abundant host among the phages with a coverage value above a given threshold was selected. The threshold was defined as a fraction of the highest coverage:

$$\text{coverage}_{\text{threshold}} = f\text{coverage}_1 \tag{7}$$

where f (fraction) is a number in the range 0.0–1.0. Note that $f = 0.0$ means considering all significant predictions, whilst $f = 1.0$ corresponds to selecting the host of the reference phage with the highest coverage (criterion 1). The optimal value of f was determined through a nested 3 fold cross-validation to avoid biased estimates of performances that would result from using the same cross validation used to select the optimal criterion. Here in turn, 3 data partitions were used as tripartite train-test set in a procedure called inner cross-validation. Within the tripartite set, 2 partitions were sequentially used as reference database for predicting the host for the left out test set using Equation (7) for a given value of f. This was repeated 3 times within each tripartite set so that all 3 partitions were used as test set and an overall performance for the given f value was calculated (see Supplementary Materials Figure S5). For each f value the average accuracy was assessed through 100 bootstrap resamplings with replacement for each inner cross validation loop. The same procedure was repeated 4 times so that each tripartite combination was analysed leading to 4 estimates of the optimal f value. The accuracy *vs.* f values curves are shown in Figure 2 for prediction of species and genus. The horizontal bars span

f values that yield at least 99% of the highest accuracy in the relative tripartite combination. Given these performance curves, an f value of 0.8 was chosen within the highest performance range, Figure 2.



(**a**) Species. (**b**) Genus.

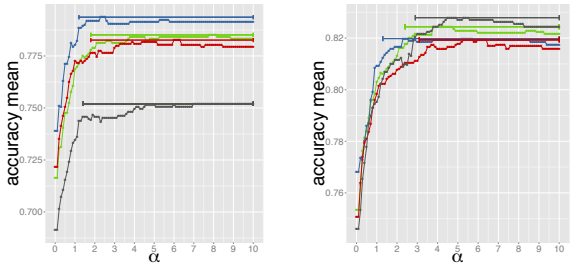
Figure 2. Accuracy *vs.* f values obtained from the 4 loops of inner cross validation. Each dot represents the averaged accuracy for species (a) and genus (b) prediction over 100 bootstrap resamplings. The bars cover the range of f values for which the accuracy is 99% the highest accuracy in the specific tripartite set.

2.4.4. Criterion 4: Summing up Normalized Coverage Values of Phages with Same Host

In the scoring method, coverage values of all significant reference phages were normalised by division by the highest coverage, coverage_1, and raised to the power of an arbitrary number, $\alpha > 0$.

$$\text{score}_i = \left(\frac{\text{coverage}_i}{\text{coverage}_1}\right)^{\alpha} \tag{8}$$

Next, scores of hits with the same host were summed up and the host was predicted as the one with the highest score. The higher the value of α, the higher the score of the first hit, the closer this method is to criterion 1. Values of α in the range 0.0–10.0 were tested. As for the criterion 3, the optimal α was determined through a nested 3 fold cross-validation setup (see Supplementary Materials Figure S5) and led to the selection of $\alpha = 6.0$ within the range that yielded the highest accuracy in the 4 tripartite train-test sets (see Figure 3).



(**a**) Species. (**b**) Genus.

Figure 3. Accuracy *vs.* α values for prediction of species (a) and genus (b) in each tripartite set. Each dot represents the averaged accuracy over 100 bootstrap resamplings. The bars cover the range of α values for which the accuracy is 99% the highest accuracy.

2.5. Programming Language and Speed of Execution

The algorithm was written in Python and Bash.

On an Intel(R) Xeon(R) CPU E5-4610 v2 @ 2.30GHz computer, using 2 cores and 10 GB RAM, HostPhinder average running time is of 61.1662 s for host species prediction and 109.622 s for genus prediction. The longer runtime for genus prediction is due the larger database used for genus predictions. These values were calculated on the evaluation set.

2.6. BLAST Evaluation

The accuracy of the HostPhinder k-mer based approach was compared to the state-of-the-art tool in bioinformatics, BLAST [51]. BLAST performance was assessed on the $\text{phages}_{\text{eval}}$ set using the $\text{phages}_{\text{train-test}}$ set to create a local nucleotide BLAST database. The host associated to the hit with the lowest E-value and secondarily highest bit score was returned as prediction.

2.7. Establishing an Evaluation Set of Predicted Prophages

The PhiSpy prophage prediction tool [52] was used to predict prophages in 2679 complete bacterial genomes collected from NCBI [53]. PhiSpy was run once on each genome resulting in a total of 7559 predicted bacterial prophages in 2074 genomes. Of these, 2796 were from bacterial species that were also included in the HostPhinder reference database. In the following, these predicted prophages will be referred to as the $\text{prophages}_{\text{species}}$ set. A total of 4639 predicted prophages were from genera that were included in the reference database of HostPhinder. They will be referred to as the $\text{prophages}_{\text{genus}}$ set.

Furthermore 261 manually verified prophages were downloaded from PhiSpy and phage_finder directories from Phantome [54] and HostPhinder prediction was tested on them.

2.8. Host Prediction of INTESTI Bacteriophage Cocktail

The Georgian George Eliava Institute of Bacteriophages, Microbiology and Virology has developed phage cocktails (mixtures of phages) since the 1950s. One of these, the INTESTI bacteriophage cocktail, claims to contain sterile filtrates of phage lysates effective against *Staphylococcus*, *Enterococcus*, *Proteus*, *Shigella*, *Salmonella*, *Escherichia coli*, and *Pseudomonas aeruginosa* for the treatment of intestinal bacterial infections. The cocktail was sequenced directly on an Illumina MiSeq platform and de novo assembled to contigs, which were further grouped into 19 draft genomes each hypothesized to represent close to complete phage genomes, and 4 smaller groups hypothesized to represent fragments of phage genomes previously described [38]. The host genus and species of each of these 23 groups was predicted by the final HostPhinder method using the 4th criterion with $\alpha = 6.0$.

3. Results

In this study, we developed and benchmarked HostPhinder, a bioinformatics tool for predicting the bacterial host species of phages. The method is based on the assumption that genetically similar phages are likely to share bacterial hosts. For performing the predictions, HostPhinder relies on a reference database in which WGS data from phages with annotated hosts have been split into k-mers. The genomes of the query phages for which the hosts should be predicted are likewise split into k-mers, and the number of co-occurring k-mers between the query phage and the phages in the reference database is used as a measure of genetic similarity.

3.1. Developing and Benchmarking the HostPhinder Method

Initial analysis on a small dataset indicated that k-mers of length 15–20 nt led to comparable predictive performances. In contrast, shorter k-mers were too unspecific and led to a lower final accuracy, while longer k-mers were too specific and led to more query phages for which no predictions at all could be made (data not shown). Based on these results and a previous study that showed

16-mers to be optimal, when using a k-mer based approach for bacterial species identification [35], 16 was chosen as the k-mer length in the following.

In the initial testing of the basic genetic similarity assumption of HostPhinder, 5 measures were evaluated for estimating the similarity of the query phage to the reference phages as described in Materials and Methods. For each measure, the query host was inferred from the host of the reference hit with the highest similarity. Table 1 shows the performance of each similarity measure in this initial testing.

Table 1. Overall performance of different similarity measures on $\text{phages}_{\text{train-test}}$.

	Score	z	frac_q	frac_d	Coverage
Species (%)	77.03 ± 0.112	77.81 ± 0.111	77.24 ± 0.111	78.43 ± 0.111	$\mathbf{78.76 \pm 0.108}$
Genus (%)	81.43 ± 0.096	82.02 ± 0.094	81.78 ± 0.094	$\mathbf{83.07 \pm 0.09}$	82.84 ± 0.092

The measures' accuracies in predicting the query phage host species of the training-test set were pairwise compared by 1000 bootstrap resamplings with replacement. Coverage performed significantly better than other measures (p-value < 0.05), apart from frac_d, which in turn did not significantly outperformed coverage. Since coverage showed the highest performance in predicting the host species, it was chosen as the measure used when further optimizing HostPhinder prediction at the species level. Next, the performance of 4 scoring methods for host selection was compared (see Material and Methods for criteria description and parameter optimization). For each selection criterion only significant hits were considered ($p_{\text{corr}} < 0.05$) and the number of queries with predictions was constant for all criteria allowing a direct comparison of criteria efficacy. Using the model parameters determined above, the 4 criteria were compared in terms of overall accuracy in a 4 fold cross-validation system. In turn, 3 of the 4 partitions were used as reference database for predicting the host for the left out test set using each criterion. This was repeated 4 times so that all 4 partitions were used as test set, and an overall performance for the given criterion was calculated. For each criterion the average and interval of confidence was assessed through 100 bootstrap resamplings with replacement for each test set and calculating the overall accuracy. Table 2 shows the overall accuracy on $\text{phages}_{\text{train-test,genus}}$ and $\text{phages}_{\text{train-test,species}}$ sets for each criterion on genus and species level, respectively. Bacterial host genera and species were not predicted for 5.8% $\text{phages}_{\text{train-test,genus}}$ and 5.6% $\text{phages}_{\text{train-test,species}}$ phages respectively.

Table 2. Average and mean standard error of the overall HostPhinder performance over 100 $\text{phages}_{\text{train-test}}$ set resamplings with replacement.

Method	Criterion 1 (First Host)	Criterion 2 (Majority Host among Top-10)	Criterion 3 (Coverage Threshold, $f = 0.8$)	Criterion 4 (Summing up Normalized Coverage Values, $\alpha = 6.0$)
Accuracy, Species (%)	78.76 ± 0.108	74.79 ± 0.102	79.1 ± 0.104	$\mathbf{79.13 \pm 0.105}$
Accuracy, Genus (%)	82.84 ± 0.092	80.41 ± 0.099	83.61 ± 0.092	$\mathbf{83.72 \pm 0.092}$

Criterion 4 with $\alpha = 6.0$ had the highest predictive value, with an accuracy of 79% and 84% for species and genus respectively, even though it only significantly outperforms criterion 2.

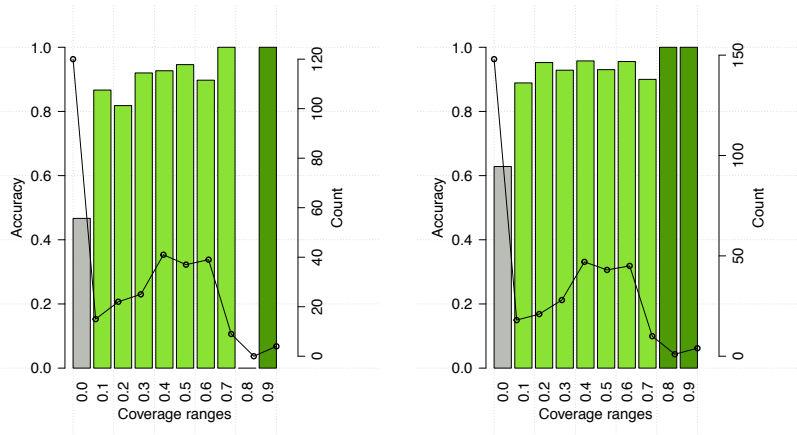
Some hosts are substantially more frequent than others in the data set. This could potentially lead to a bias in the prediction, and a subsequent sub-optimal predictive performance. To investigate this, modified versions of criteria 2–4 were tested, where the sequences in the reference database were

clustered according to Hobohm 1 algorithm [49], and only the highest scoring element within one cluster was used in the prediction schema. This did not, however, improve the performance.

Based on the above benchmarking procedures, the final method called HostPhinder was developed. The reference database was generated by splitting all phage genomes in the entire phage set into 16-mers using a step-size of 1. After searching through the database, HostPhinder examines the coverage measure and creates a hits list, *i.e.*, phages significantly similar to the query. The final host species and genus is given according to criterion 4 with an $\alpha = 6.0$. HostPhinder is freely available as a web server [1] and as a Docker image [2].

3.2. Evaluating HostPhinder's Performance on Complete and Partial Genomes

HostPhinder was evaluated on the $\text{phages}_{\text{eval,genus}}$ and $\text{phages}_{\text{eval,species}}$ sets containing phages from public databases. HostPhinder was able to correctly predict the bacterial host species and genera of 74.24% $\pm$ 0.270% and 81.39% $\pm$ 0.206% of the phages respectively. In the evaluation set, 4.0% (3.44%) of the phages could not be matched to any phage in the database when predicting on species (genus) level. We speculated that the accuracy of the HostPhinder method is depending on the coverage value of its prediction. That is, the higher the coverage value, the higher the accuracy. To quantify if this is indeed the case, we show in Figure 4 the accuracy on the evaluation set at different intervals of the coverage value. No hit appeared to have range 0.8 < coverage $\leq$ 0.9 for species. For species as well as genus level, it can be seen that predictions based on a coverage value below 0.1 are only correct for 47% (species) and 63% (genus) of the phages. At the other end of the scale, predictions based on a coverage value above 0.7 (species) and 0.8 (genus) are correct in all instances.

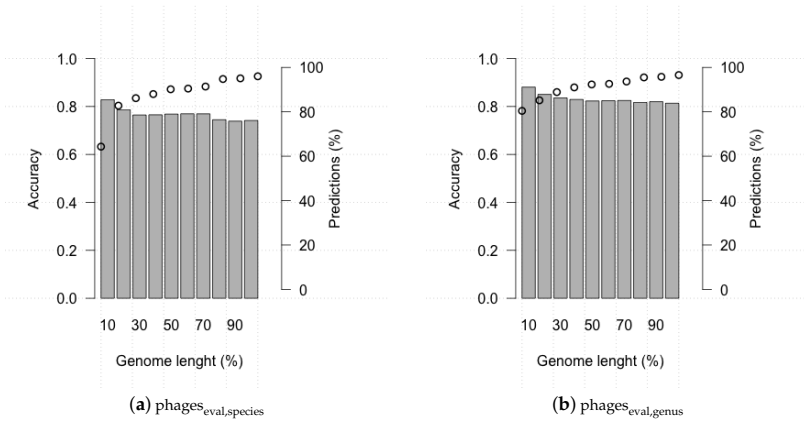


(**a**) $\text{phages}_{\text{eval,species}}$ (**b**) $\text{phages}_{\text{eval,genus}}$

Figure 4. HostPhinder's accuracy (bar) and prediction counts (line) on $\text{phages}_{\text{eval}}$ at different coverage ranges. The values displayed on the x axis are the lower limit of that range. With exception of the last bin which includes all entries with coverage >0.9, all ranges are right-closed with upper limit $x + 0.1$. Poorly reliable results are in grey, while reliable and highly reliable results are in green and dark green respectively. Results on HostPhinder's web server [1] are displayed using the same colour code.

Assembly of metagenomic samples often do not results in entire phage genomes. To assess how the completeness of a phage genome affects HostPhinder performance, we ran the tool on the evaluation set where each genome was gradually reduced by 10%, 20%, ... ,90% of its total length.

Figure 5 shows the accuracy and the number of predictions for each percentage of genome length. HostPhinder maintained the prediction accuracy but made gradually fewer predictions as the fraction of genome given as query is decreased.



(**a**) $\text{phages}_{\text{eval,species}}$ (**b**) $\text{phages}_{\text{eval,genus}}$

Figure 5. HostPhinder's accuracy (bar) and percentages of predictions (dots) on $\text{phages}_{\text{eval}}$ at different percentages of genome length from 10% to 100% of total genome length.

Generally, HostPhinder returned predictions at 10% genome length for those genomes which prediction at complete genome length had a higher coverage. The average coverage for predictions made at complete genome length but not at 10% genome length was 0.023, while the average coverage for commonly predicted was 0.36.

We next examined if HostPhinder always correctly predicted particular host species or genera (Table 3). Only hosts occurring at least 3 times in the $\text{phages}_{\text{eval}}$ set are listed. All phages in the $\text{phages}_{\text{eval}}$ set that target these hosts listed in Table 3 were correctly predicted. Additionally, none of these hosts were erroneously predicted as targets of other phages.

Table 3. List of host species (left) and genera (right), which HostPhinder predicts correctly.

Species	**Representation in $\text{phages}_{\text{train-test,species}}$**	**Genus**	**Representation in $\text{phages}_{\text{train-test,genus}}$**
Enterococcus faecalis	15	*Acinetobacter*	16
Listeria monocytogenes	21	*Listeria*	26
Propionibacterium acnes	21	*Propionibacterium*	24
Vibrio cholerae	35	*Streptococcus*	39
		Streptomyces	11
		Thermus	5

HostPhinder also worked effectively for predicting the host of phages, which according to the initial clustering were of different types; in fact in the HostPhinder dataset there are 14 different types of *Enterococcus faecalis* phages, 13 types of *Listeria monocytogenes* phages and 21 types of *Vibrio cholerae* phages and all phages known to infect these host have been correctly predicted, see Table 3.

Figures 6 and 7 show right and wrong predictions for species and genera respectively. To ease comprehension of the plots, hosts were grouped by phyla, which are displayed on the left side of the figures. Rows are alternatively shaded and column names are enhanced with the same

colour of the phylum of belonging. The heatmaps are read from right to left and then downwards; expressely, the phage related to the host identified by the row name, on the right, was predicted (red intensity of the cell) to infect the host identified by the column name in the lower part of the figure. As an example, *Alteromonas macleodii* phages, the row encompassed in a blue horizontal box in Figure 6, occurred four times in the $\text{phages}_{\text{eval,species}}$ set, as indicated by the number within parenthesis beside the host name, and all of them were wrongly predicted to be *S. aureus* phages (vertical blue box) as indicated by the intense red colour of the square in the intersection between the two blue boxes; of note, there were 69 *S. aureus* phages in the $\text{phages}_{\text{train-test,species}}$ data set and no *Alteromonas macleodii* phages.

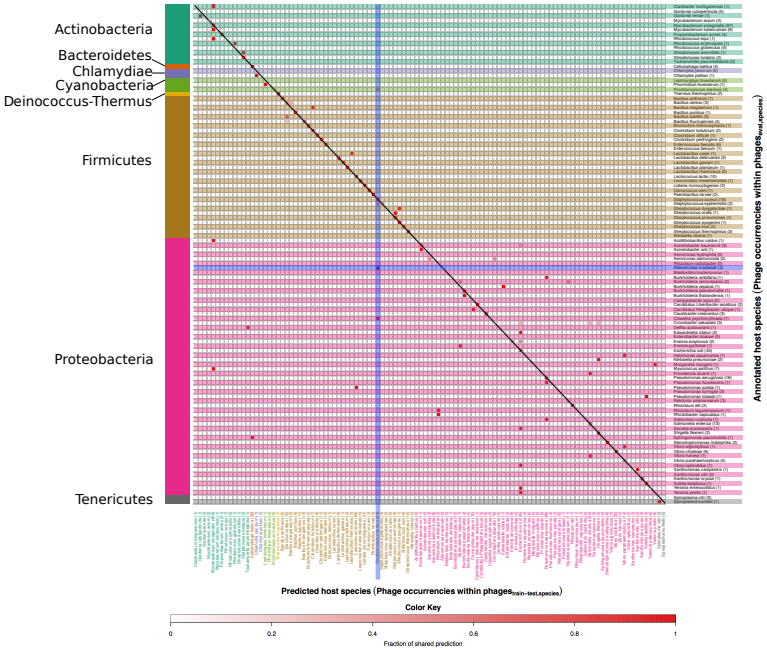


Figure 6. Heatmap of annotated *vs.* predicted host species in the $\text{phages}_{\text{eval,species}}$ set. In this figure correct as well as mispredicted host species can be seen. Annotated host species are listed along the *y* axis, while predicted ones are on the *x* axis. The number after each species on the *y* axis and the *x* axis also indicate the occurrences of phages in the $\text{phage}_{\text{eval,species}}$ and in the $\text{phages}_{\text{train-test}}$ respectively. Host species are grouped according to the respective phylum, which are indicated on the left side of the figure. The colour scale indicates the fraction of phages predicted as targeting a particular host and goes from white, no phages, to red, 100% of the phages. Accordingly, the colour itself is not an indicator of correctness of the prediction, and red colours along the diagonal represent correct predictions.

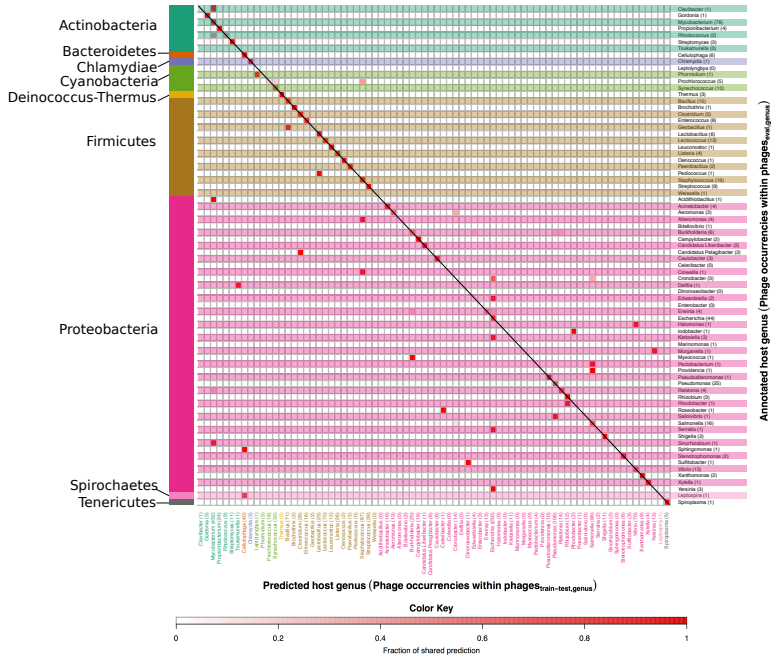


Figure 7. Heatmap of annotated *vs.* predicted host genera in the phages$_{\text{eval,genera}}$ set. In this figure correct as well as mispredicted host genera can be seen. Annotated host genera are listed along the y axis, while predicted ones are on the x axis. The number after each genus on the y axis and the x axis indicate the number of occurrences of phages in the phage$_{\text{eval,genus}}$ and phages$_{\text{train-test,genus}}$ respectively. Host genera are grouped according to the respective phylum, which are indicated on the left side of the figure. The colour scale indicates the fraction of phages predicted as targeting a particular host and goes from white, no phages, to intense red, 100% of the phages. Accordingly, the colour is in itself not an indicator of correctness of the prediction, and red colours along the diagonal represent correct predictions.

At species level, phages with mispredicted hosts are often predicted to target a host of the same genus as the annotated host (see small deviations from the diagonal in Figure 6). As examples, the 3 phages annotated to target *Bacillus subtilis* are predicted to target either *B. subtilis* or *Bacillus cereus*. For some phages the mispredicted host is, however, of an entirely different genus, e.g., the phage annotated to target *Yersinia enterocolitica* and the phage annotated to target *Yersinia pestis* are both predicted to target *E. coli*. For species as well as genera there is a tendency that phages with mispredicted hosts are predicted to target the most frequent hosts in the phages$_{\text{train-test}}$ set, e.g., *E. coli* and *Mycobacterium smegmatis* on species level and *Escherichia* and *Mycobacterium* on genus level. What is important to note is that inaccurate predictions were finding related hosts. For example, imprecise predictions of phages infecting *Proteobacteria* (the ones within the brown region) were still falling within the phylum of *Proteobacteria*. This indicates a relatedness in terms of genome sequence among phages infecting different hosts belonging to the same phylum.

3.3. Comparing HostPhinder to BLAST

Next, the HostPhinder performance on $phages_{eval}$ was compared to BLAST. Table 4 summarises the results.

Table 4. HostPhinder and BLAST performance comparison on the $phages_{eval}$ set.

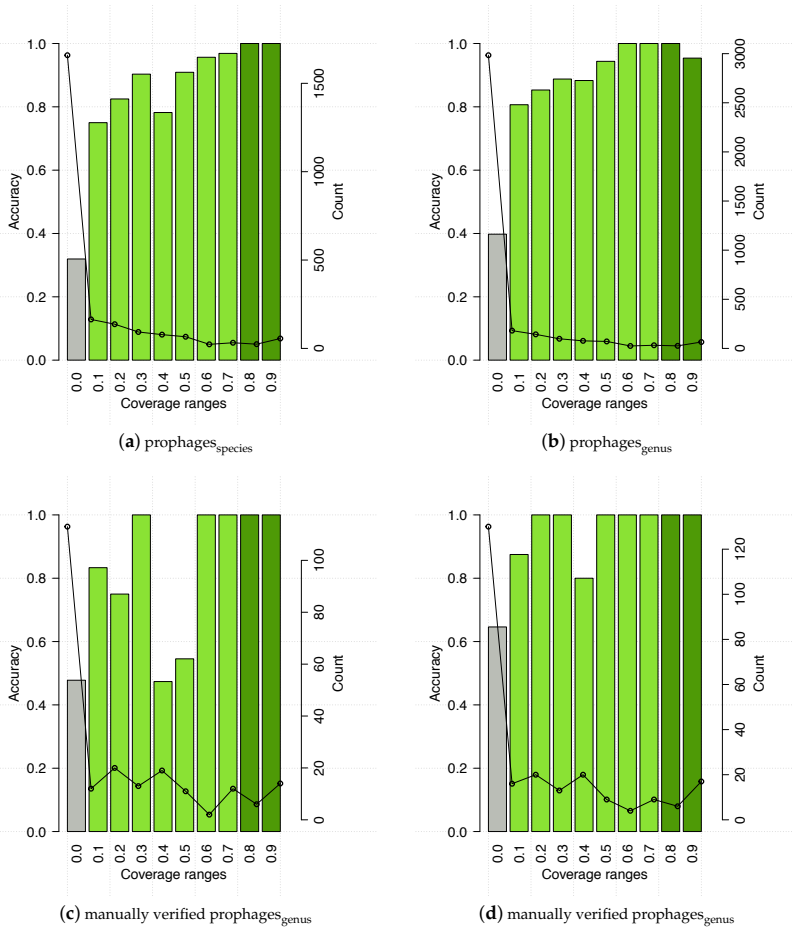
	BLAST	**HostPhinder**
No. of predictions, training on $phages_{train\text{-}test,genus}$	90%	97%
No. of predictions, training on $phages_{train\text{-}test,species}$	91%	96%
Accuracy on common predictions (GENERA) (%)	84.66 ± 0.188	85.13 ± 0.176
Accuracy on common predictions (SPECIES) (%)	76.92 ± 0.252	78.69 ± 0.237

HostPhinder was able to make host predictions for more phages than the BLAST-based method. For the phages that both methods were able to make a prediction for, HostPhinder outperformed BLAST on both genus and species level. The observed better performance of HostPhinder on species level is significant ($p < 0.05$). HostPhinder correctly predicted 25% among 24 (genera) and 10% among 20 (species) predictions not covered by BLAST. Moreover when inferring the host genus of a phage for which HostPhinder gave no prediction, BLAST match to the most closely related phage resulted in the wrong prediction.

3.4. HostPhinder's Performance on Predicted Prophages and Establishment of Confidence Threshold

To further evaluate the performance of HostPhinder and to establish a confidence threshold for the predictive value, we examined if HostPhinder was able to identify the bacterial hosts of predicted prophages on the premise that prophages are phages that have at one point infected the host that they are currently found in. The predicted prophages provide a dataset diverse enough to define a reliability threshold that can be generalized and applied to previously unseen data. For this purpose, we predicted prophages in 2679 bacterial genomes using PhiSpy [52]. Without any threshold value set, HostPhinder was able to correctly predict approximately 45% and 47% of the species and genus respectively. The accuracy was calculated over the number of phages that HostPhinder was able to make a prediction for.

As for $phages_{eval}$, the results on PhiSpy predicted prophages were binned into coverage ranges (Figure 8, upper panels). The accuracy pattern for prophages generally resembled the one for the evaluation set, *i.e.*, it had low accuracy for coverage ≤ 1, and 100% accuracy above a certain threshold, which in this case is 0.8 for species. There is an unexpected drop in accuracy for coverage values >0.9 (genus), which a bootstrap analysis proved non significant ($p > 0.05$). To further cofirm the thresholds, we ran HostPhinder on 261 manually verified prophages, downloaded from PhAnToMe.org, which resulted in 63.57 % $\pm$ 0.356 % and 78.69 % $\pm$ 0.262 % prediction accuracy of species and genus respectively. Accuracy distribution for this dataset among different coverage ranges can be seen in Figure 8, lower panels. Based on observations $phages_{eval}$ and on prophages, HostPhinder considers trustable results with coverage value higher than 0.1, and it applies a conservative threshold of 0.8 to distinguish highly trustable results.



(**a**) prophages$_{species}$ (**b**) prophages$_{genus}$

(**c**) manually verified prophages$_{genus}$ (**d**) manually verified prophages$_{genus}$

Figure 8. HostPhinder's accuracy (bar) and prediction counts (line) on prophages predicted by PhySpy, upper panels, and manually verified prophages, lower panels, at different coverage ranges. The values displayed on the x axis are the lower limit of that range. With exception of the last bin which includes all entries with coverage >0.9, all ranges are right-closed with upper limit $x + 0.1$. Poorly reliable results are in grey, while reliable and highly reliable results are in green and dark green respectively.

3.5. Host Analysis of Phages from Therapeutic Phage Cocktail from the Georgian George Eliava Institute

In a recent study, we examined the content of an INTESTI bacteriophage cocktail from the Georgian George Eliava Institute. According to the packing, the cocktail is effective against *Staphylococcus*, *Enterococcus*, *Proteus*, *Shigella*, *Salmonella*, *Escherichia coli*, and *Pseudomonas aeruginosa* infections [38]. A total of 19 phage draft genomes were identified that were hypothesized to represent close to complete phage genomes. An additional set of four sequences represented fragments of phage

genomes. Here, we used HostPhinder in an attempt to predict host genera and species of these phage draft genomes and fragments. Table 5 provides an overview.

Table 5. Overview of the results of HostPhinder predicting the hosts of 19 phage draft genomes (name starts with a "D" and *Proteus*) and 4 phage genome fragments (name starts with an "F") from the INTESTI phage cocktail.

Draft ID	Genus	Species	Coverage
D1	*Staphylococcus*	*Staphylococcus aureus*	1.000
D13	*Salmonella*	*Salmonella enterica*	0.840
D5	*Escherichia*	*Escherichia coli*	0.740
D3	*Pseudomonas*	*Pseudomonas aeruginosa*	0.690
D11	*Salmonella*	*Salmonella enterica*	0.610
D10	*Escherichia*	*Escherichia coli*	0.500
D14	*Escherichia*	*Escherichia coli*	0.490
D17	*Escherichia*	*Escherichia coli*	0.460
D9	*Escherichia*	*Escherichia coli*	0.450
D15	*Escherichia*	*Escherichia coli*	0.450
D16	*Sodalis*	*Sodalis glossinidius*	0.430
D4	*Escherichia*	*Escherichia coli*	0.420
D18	*Salmonella*	*Salmonella enterica*	0.380
D8	*Escherichia*	*Shigella flexneri*	0.300
F1	*Pseudomonas*	*Pseudomonas aeruginosa*	0.270
D2	*Escherichia*	*Escherichia coli*	0.250
D7	*Enterococcus*	*Enterococcus faecalis*	0.230
D12	*Enterococcus*	*Enterococcus faecium*	0.180
F2	*Salmonella*	*Salmonella enterica*	0.078
F4	*Escherichia*	*Escherichia coli*	0.014
D6	*Enterococcus*	*Enterococcus faecalis*	0.011
F3	*Escherichia*	*Escherichia coli*	0.011
Proteus	*Escherichia*	*Escherichia coli*	0.003

For six of the seven bacterial targets of the cocktail, HostPhinder predicted at least one phage targeting this type of bacteria. The only bacterium that was not predicted among the hosts was *Proteus*. Instead, the phage that was experimentally found to infect *Proteus* [38], was predicted as an *E. coli* phage with a coverage of 0.0026. This is not surprising, as the HostPhinder database contains no examples of *Proteus* phages. A *Sodalis glossinidius* was predicted, not corresponding to any of the anticipated targets. This bacterium is an endosymbiont of the tsetse fly [50] and its prediction was based on a coverage value of 0.43, where predictions with coverages above 0.2 have approximately 80% chance of being correct (see Figures 4 and 8). The predicted hosts of the 4 phage fragments were generally based on a lower coverage than the 19 phage draft genomes, indicating that these predictions are less certain.

4. Discussion

In the present study, we developed a fast and simple method for prediction of phage hosts. Other studies have previously focused on the identification of phage-host pairs. Experimental methods examining phage-host interactions include mining viral signals from SAG (single amplified genomes) datasets; microfluidic digital PCR and phageFISH [55]. Recently, M. Martínez-García *et al.* combined single-cell genomics and microarrays technology to assign viruses to hosts depending on hybridization

allowing for discovery of new virus-host pairs directly on a metagenomic samples without requiring cultivation or relying on genomic information [56]. In another study, Roux *et al.* developed a bioinformatics tool VirSorter [57], which was able to identify more than 12,000 virus-host linkages from publicly available bacterial and archeal genomes. In their study they analysed the virus-host adaptation in compositions in terms of mono- di- tri- tetra-nucleotide frequency and codon usage [58] showing the strongest signal of adaptation to host genome given by tetranucleotide frequency (TNF). A further classification method for phage host prediction, MGTAXA was developed by Williamson *et al.* in their metagenomic study of the marine microbe in the Indian Ocean [59]. MGTAXA links viral sequences to the highest scoring host taxonomic model based on polynucleotide genome composition similarity between phage and bacterial genomes. The software is not conveniently available anymore (as of December 2015) and we therefore could not compare its performance to HostPhinder's. Finally, a recent publication by Edwards *et al.* reviewed the predictive power of several computational tools for predicting the host of a given phage based on genome information [60]. The authors highlighted the importance of such tools for the characterization of uncoltured virus from metagenomes, and found that homology-based approaches had the strongest signals for predicting phage-host interactions.

HostPhinder bases its predictions on co-occurring k-mers between the query phage genome and the genomes of reference phages with known hosts. Kmer-based approaches have recently been implemented for genome assembly [61], fast classification [62,63] and annotation [64] of metagenomes. Considering the highly mosaic structure of phage genomes, one of the advantages of using k-mers for phage host predictions is that the exact order of genetic elements does not influence the outcome, only their presence or absence.

On an independent evaluation set, HostPhinder was found to perform well, when predicting the hosts of phages currently found in public databases. A remarkable 74% accuracy for the host species and 81% for the host genus were obtained. Some hosts were consistently easier to predict than others. This was for example the case for *P. acnes*, where the host of all annotated *P. acnes* phages in the evaluation set were correctly predicted, while no non-*P. acnes* phages were erroneously predicted as such. The observation is in concordance with previous studies showing that *P. acnes* phages constitute a homogenous group, sharing 85% nucleotide sequence and having similar genome length [65,66]. Furthermore the examined *P. acnes* phages were not able to infect other members of the *Propionibacterium* genus [65,67]. For many of the mispredicted hosts of HostPhinder, the genus of the annotated and predicted host was the same, which might be considered concurrent with the ability of some phages to infect more than one species within a genus. Examples of such broad host range phages are *Salmonella* Phage Felix O1 [68], Mycobacteriophage D29 [69] and *Yersinia* Phage PY100 [70]. It is hence possible that the mispredicted phages are polyvalent, *i.e.*, capable of infecting more than one bacterial species. Alternatively they may represent actual misprediction by HostPhinder caused by closely related phages targeting different host species. In some cases, the host predicted by HostPhinder did not even belong to the same genus as the annotated host, e.g., the three *Yersinia* phages were all predicted to infect *Escherichia* with coverage values that indicate a reliable result, namely 0.57, 0.6 and 0.13. Indeed the genome sequence of the *Y. pestis* phage phiA1122 has been found to be closely related to coliphage T7, sharing 89% nucleotide identity [71]. Despite this high nucleotide identity, PhiA1122 is not able to infect *E. coli*, and has even been used by the Center for Disease Control and Prevention of the United States as a diagnostic agent to identify *Y. pestis* [72].

When applying HostPhinder to phage draft genomes and fragments from the INTESTI phage cocktail, the predicted hosts corresponded well with the advertised targets of the cocktail. One phage draft genome was, however, predicted to target *Sodalis glossinidius*, an endosymbiont of the tsetse fly. Excluding the remote possibility that phages targeting this bacterium has been added to the cocktail, it is likely that the HostPhinder prediction is incorrect or that the phage is able to infect *S. glossinidius* as well as one of the targets of the cocktail. A study by Ho-Won and Kyoung-Ho Kim has shown close relation in comparative genomic and phylogenetic analyses between EP23, a phage that infects

E. coli and *Shigella sonnei* and, SO-1, which infects *S. glossinidius* [73]. It was, however, not examined if the phages were able to cross-infect the hosts.

Many phages have a very narrow host range and only target specific strains within a particular species. This feature has been used extensively previously, when typing, e.g., *S. enterica* [74] and *S. aureus* [75]. HostPhinder is not able to perform predictions beyond species level, partly due to the hosts of most phages in the public databases not being annotated beyond this. Further, to perform predictions down to specific strains of bacteria more factors than the mere genome resemblance would likely have to be taken into account, e.g., by examining the receptor binding proteins, identifying the number of restriction sites in the phage genomes or analysing the CRISPR regions of the host genome.

Another limitation to the performance of HostPhinder is the accuracy of the breadth of annotated host(s) of the references phages. Most of the reference phages had only one annotated host, although many examples exist of phages that are able to infect closely or even distantly related bacteria [76–78]. Further, the performance of HostPhinder depends on the size and completeness of the underlying database. As an example, at the time of compiling the database for this study, no *Proteus* phage genomes were available in public databases. Hence it is inherently impossible for the HostPhinder method to predict any query phage as a *Proteus* phage. Indeed, HostPhinder predicted an experimentally identified *Proteus* phage from the INTESTI phage cocktail as an *E. coli* phage, albeit based on a coverage value of 0.003 indicating that the prediction was not reliable. Carson *et al.* demonstrated the capability of a coli-proteus phage isolated from a Russian cocktail of equally eradicating *E. coli* and *Proteus mirabilis* biofilms [79], evincing the potential of some phages to infect both species. As more phage genomes become available, we will update HostPhinder database to ensure its continued high performance.

Despite the limitations in HostPhinder, we envision that the tool will be useful for narrowing down the list of potential hosts. With the growing availability of metagenome samples, new approaches are necessary to firstly identify phages and secondly, determine their host. Thanks to its capability of promptly identifying potential phage-host interactions, the HostPhinder tool has potential applications in ecology, human gut microbiocenosis studies, and other viral metagenomics analyses, where there is need to shed light on the nature of phages.

The current of HostPhinder is very simple, only taking into account genomic information about the phage. Further development of the tool will expand this, taking the genome of the host into account, which we expect will enable us to make predictions beyond host species level.

5. Conclusions

The current antibiotics resistance crisis warrants new ways to combat bacterial infections. For decades, phage therapy has been used for this purpose in countries belonging to the former Eastern Bloc, and to ensure transfer of the technology to the West, it is important to establish a pool of well-characterized phages. The presented HostPhinder method provides the phage community with an easy-to-use tool for predicting the host genus and species of query phages, usable when searching for phages with appropriate host specificity and for correlating phages and hosts in ecological and metagenomic studies. HostPhinder is freely available as a web server [1] and as a Docker image [2].

Acknowledgments: This work was supported by the Center for Genomic Epidemiology at the Technical University of Denmark and funded by grant 09-067103/DSF from the Danish Council for Strategic Research, grant nr. 14-3056 from Oticon Fonden and grant 14-70-0955 from Otto Moensted Fonden.

Author Contributions: Mette Voldby Larsen conceived the method; Ole Lund wrote the script to read the k-mers; Kortine Annina Kleinheinz developed the preliminary version of the method; Morten Nielsen designed the method optimization; Julia Villarroel downloaded whole genome sequence data, performed method optimization, analysed the data and finalized the method; Henrike Zschach predicted the prophages; Vanessa Isabell Jurtz designed the Hobohm experiments and set up HostPhinder web server; Julia Villarroel built HostPhinder Docker image; Julia Villarroel, Mette Voldby Larsen and Morten Nielsen wrote the paper. All authors contributed in reviewing the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. HostPhinder web service. Available online: http://cge.cbs.dtu.dk/services/HostPhinder (accessed on 1 April 2016).
2. HostPhinder Docker image. Available online: https://registry.hub.docker.com/u/julvi/hostphinder (accessed on 1 April 2016).
3. Kapi, A. The evolving threat of antimicrobial resistance: Options for action. *Indian J. Med. Res.* **2014**, *139*, 182–183.
4. WHO. *Antimicrobial Resistance: Global Report on Surveillance*; World Health Organization: Geneva, Switzerland, 2014.
5. Harper, D.; Anderson, J.; Enright, M. Phage therapy: Delivering on the promise. *Ther. Deliv.* **2011**, *2*, 935–947.
6. Kutateladze, M.; Adamia, R. Bacteriophages as potential new therapeutics to replace or supplement antibiotics. *Trends Biotechnol.* **2010**, *28*, 591–595.
7. Kutateladze, M.; Adamia, R. Phage therapy experience at the Eliava Institute. *Méd. Mal. Infect.* **2008**, *38*, 426–430.
8. Miedzybrodzki, R.; Borysowski, J.; Weber-Dabrowska, B.; Fortuna, W.; Letkiewicz, S.; Szufnarowski, K.; Pawelczyk, Z.; Rogóz, P.; Klak, M.; Wojtasik, E.; *et al*. Charpter 3—Clinical aspects of phage therapy. *Adv. Virus Res.* **2012**, *83*, 73–121.
9. Weber-Dąbrowska, B.; Mulczyk, M.; Górski, A. Bacteriophage therapy of bacterial infections: An update of our institute's experience. In *Inflammation*; Springer: Netherlands, 2001; pp. 201–209.
10. Biswas, B.; Adhya, S.; Washart, P.; Paul, B.; Trostel, A.N.; Powell, B.; Carlton, R.; Merril, C.R. Bacteriophage therapy rescues mice bacteremic from a clinical isolate of vancomycin-resistant *Enterococcus faecium*. *Infect. Immun.* **2002**, *70*, 204–210.
11. Capparelli, R.; Parlato, M.; Borriello, G.; Salvatore, P.; Iannelli, D. Experimental phage therapy against *Staphylococcus aureus* in mice. *Antimicrob. Agents Chemother.* **2007**, *51*, 2765–2773.
12. Smith, H.W.; Huggins, M. Successful treatment of experimental *Escherichia coli* infections in mice using phage: Its general superiority over antibiotics. *J. Gen. Microbiol.* **1982**, *128*, 307–318.
13. Wright, A.; Hawkins, C.; Änggård, E.; Harper, D. A controlled clinical trial of a therapeutic bacteriophage preparation in chronic otitis due to antibiotic-resistant *Pseudomonas aeruginosa*; A preliminary report of efficacy. *Clin. Otolaryngol.* **2009**, *34*, 349–357.
14. Matsuzaki, S.; Uchiyama, J.; Takemura-Uchiyama, I.; Daibata, M. Perspective: The age of the phage. *Nature* **2014**, *509*, doi:10.1038/509S9a.
15. Reardon, S. Phage therapy gets revitalized. *Nature* **2014**, *510*, doi:10.1038/510015a.
16. Sulakvelidze, A. Using lytic bacteriophages to eliminate or significantly reduce contamination of food by foodborne bacterial pathogens. *J. Sci. Food Agric.* **2013**, *93*, 3137–3146.
17. Guenther, S.; Huwyler, D.; Richard, S.; Loessner, M.J. Virulent bacteriophage for efficient biocontrol of *Listeria monocytogenes* in ready-to-eat foods. *Appl. Environ. Microbiol.* **2009**, *75*, 93–100.
18. Carrillo, C.L.; Atterbury, R.; El-Shibiny, A.; Connerton, P.; Dillon, E.; Scott, A.; Connerton, I. Bacteriophage therapy to reduce *Campylobacter jejuni* colonization of broiler chickens. *Appl. Environ. Microbiol.* **2005**, *71*, 6554–6563.
19. McLean, S.K.; Dunn, L.A.; Palombo, E.A. Phage inhibition of *Escherichia coli* in ultrahigh-temperature-treated and raw milk. *Foodborne Pathog. Dis.* **2013**, *10*, 956–962.
20. Stern, A.; Sorek, R. The phage-host arms race: Shaping the evolution of microbes. *Bioessays* **2011**, *33*, 43–51.
21. Deveau, H.; Garneau, J.E.; Moineau, S. CRISPR/Cas system and its role in phage-bacteria interactions. *Annu. Rev. Microbiol.* **2010**, *64*, 475–493.
22. Fineran, P.C.; Blower, T.R.; Foulds, I.J.; Humphreys, D.P.; Lilley, K.S.; Salmond, G.P. The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 894–899.

23. Carbone, A. Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J. Mol. Evol.* **2008**, *66*, 210–223.
24. Blower, T.R.; Evans, T.J.; Przybilski, R.; Fineran, P.C.; Salmond, G.P. Viral evasion of a bacterial suicide system by RNA-based molecular mimicry enables infectious altruism. *PLoS Genet.* **2012**, *8*, e1003023.
25. Labrie, S.J.; Samson, J.E.; Moineau, S. Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* **2010**, *8*, 317–327.
26. Weitz, J.S.; Hartman, H.; Levin, S.A. Coevolutionary arms races between bacteria and bacteriophage. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 9535–9540.
27. Duffy, S.; Turner, P.E.; Burch, C.L. Pleiotropic costs of niche expansion in the RNA bacteriophage Φ6. *Genetics* **2006**, *172*, 751–757.
28. Amarillas, L.; Cháidez-Quiroz, C.; Sañudo-Barajas, A.; León-Félix, J. Complete genome sequence of a polyvalent bacteriophage, phiKP26, active on *Salmonella* and *Escherichia coli*. *Arch. Virol.* **2013**, *158*, 2395–2398.
29. Loessner, M.J.; Neugirg, E.; Zink, R.; Scherer, S. Isolation, classification and molecular characterization of bacteriophages for Enterobacter species. *J. Gen. Microbiol.* **1993**, *139*, 2627–2633.
30. Koskella, B.; Meaden, S. Understanding bacteriophage specificity in natural microbial communities. *Viruses* **2013**, *5*, 806–823.
31. Casjens, S.R. Diversity among the tailed-bacteriophages that infect the Enterobacteriaceae. *Res. Microbiol.* **2008**, *159*, 340–348.
32. Rohwer, F.; Edwards, R. The Phage Proteomic Tree: A genome-based taxonomy for phage. *J. Bacteriol.* **2002**, *184*, 4529–4535.
33. Jacobs-Sera, D.; Marinelli, L.J.; Bowman, C.; Broussard, G.W.; Bustamante, C.G.; Boyle, M.M.; Petrova, Z.O.; Dedrick, R.M.; Pope, W.H.; Advancing, S.E.A.P.H.; *et al*. On the nature of mycobacteriophage diversity and host preference. *Virology* **2012**, *434*, 187–201.
34. Woese, C.R.; Fox, G.E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5088–5090.
35. Larsen, M.V.; Cosentino, S.; Lukjancenko, O.; Saputra, D.; Rasmussen, S.; Hasman, H.; Sicheritz-Pontén, T.; Aarestrup, F.M.; Ussery, D.W.; Lund, O. Benchmarking of methods for genomic taxonomy. *J. Clin. Microbiol.* **2014**, *52*, 1529–1539.
36. Hendrix, R.W. Bacteriophage genomics. *Curr. Opin. Microbiol.* **2003**, *6*, 506–511.
37. Lawrence, J.G.; Hatfull, G.F.; Hendrix, R.W. Imbroglios of viral taxonomy: Genetic exchange and failings of phenetic approaches. *J. Bacteriol.* **2002**, *184*, 4891–4905.
38. Zschach, H.; Joensen, K.G.; Lindhard, B.; Lund, O.; Goderdzishvili, M.; Chkonia, I.; Jgenti, G.; Kvatadze, N.; Alavidze, Z.; Kutter, E.M.; *et al*. What can we learn from a metagenomic analysis of a Georgian bacteriophage cocktail? *Viruses* **2015**, *7*, 6570–6589.
39. Nielsen, H.B.; Almeida, M.; Juncker, A.S.; Rasmussen, S.; Li, J.; Sunagawa, S.; Plichta, D.R.; Gautier, L.; Pedersen, A.G.; le Chatelier, E.; *et al*. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **2014**, *32*, 822–828.
40. Phages.ids - VBI mirrors page. Available online: http://mirrors.vbi.vt.edu/mirrors/ftp.ncbi.nih.gov/genomes/IDS/Phages.ids (accessed on 1 April 2016).
41. NCBI viral Genome Resource. Available online: http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi (accessed on 1 April 2016).
42. EMBL EBI phage genomes list. Available online: http://www.ebi.ac.uk/genomes/phage.html (accessed on 1 April 2016).
43. phagesdb for Mycobacteriophages. Available online: http://phagesdb.org/ (accessed on 1 April 2016).
44. phagesdb for Arthrobacter. Available online: http://arthrobacter.phagesdb.org/ (accessed on 1 April 2016).
45. phagesdb for Bacillus. Available online: http://bacillus.phagesdb.org/ (accessed on 1 April 2016).
46. phagesdb for Streptomyces. Available online: http://streptomyces.phagesdb.org/ (accessed on 1 April 2016).
47. Euzéby, J.P. List of Bacterial Names with Standing in Nomenclature: A folder available on the Internet. *Int. J. Syst. Bacteriol.* **1997**, *47*, 590–592.
48. HostPhinder Github repository. Available online: https://github.com/julvi/HostPhinder (accessed on 1 April 2016).
49. Hobohm, U.; Scharf, M.; Schneider, R.; Sander, C. Selection of representative protein data sets. *Protein Sci.* **1992**, *1*, 409–417.

50. Bonferroni, C.E. *Teoria Statistica Delle Classi e Calcolo Delle Probabilita*; Libreria Internazionale Seeber: Firenze, Italy, 1936.
51. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.
52. Akhter, S.; Aziz, R.K.; Edwards, R.A. PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Res.* **2012**, *40*, doi:10.1093/nar/gks406.
53. NCBI complete bacterial genomes. Available online: ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/ (accessed on 1 April 2016).
54. Phantome manually verified prophages, dating 14 March 2012. Available online: http://www.phantome.org/Downloads/Prophages/PhiSpy/Manually_Verified/ (accessed on 1 April 2016).
55. Dang, V.T.; Sullivan, M.B. Emerging methods to study bacteriophage infection at the single-cell level. *Front. Microbiol.* **2014**, *5*, doi:10.3389/fmicb.2014.00724.
56. Martínez-García, M.; Santos, F.; Moreno-Paz, M.; Parro, V.; Antón, J. Unveiling viral-host interactions within the 'microbial dark matter'. *Nat. Commun.* **2014**, *5*, doi:10.1038/ncomms5542.
57. Roux, S.; Enault, F.; Hurwitz, B.L.; Sullivan, M.B. VirSorter: Mining viral signal from microbial genomic data. *PeerJ* **2015**, *3*, doi:10.7717/peerj.985.
58. Roux, S.; Hallam, S.J.; Woyke, T.; Sullivan, M.B. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife* **2015**, *4*, doi:10.7554/eLife.08490.
59. Williamson, S.J.; Allen, L.Z.; Lorenzi, H.A.; Fadrosh, D.W.; Brami, D.; Thiagarajan, M.; McCrow, J.P.; Tovchigrechko, A.; Yooseph, S.; Venter, J.C. Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS ONE* **2012**, *7*, e42047.
60. Edwards, R.A.; McNair, K.; Faust, K.; Raes, J.; Dutilh, B.E. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.* **2016**, *40*, 258–272.
61. Marçais, G.; Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **2011**, *27*, 764–770.
62. Kawulok, J.; Deorowicz, S. CoMeta: Classification of metagenomes using k-mers. *PLoS ONE* **2015**, *10*, e0121453.
63. Wood, D.E.; Salzberg, S.L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **2014**, *15*, doi:10.1186/gb-2014-15-3-r46.
64. Edwards, R.A.; Olson, R.; Disz, T.; Pusch, G.D.; Vonstein, V.; Stevens, R.; Overbeek, R. Real Time Metagenomics: Using k-mers to annotate metagenomes. *Bioinformatics* **2012**, *28*, 3316–3317.
65. Marinelli, L.J.; Fitz-Gibbon, S.; Hayes, C.; Bowman, C.; Inkeles, M.; Loncaric, A.; Russell, D.A.; Jacobs-Sera, D.; Cokus, S.; Pellegrini, M.; *et al*. *Propionibacterium acnes* bacteriophages display limited genetic diversity and broad killing activity against bacterial skin isolates. *MBio* **2012**, *3*, doi:10.1128/mBio.00279-12 .
66. Liu, J.; Yan, R.; Zhong, Q.; Ngo, S.; Bangayan, N.J.; Nguyen, L.; Lui, T.; Liu, M.; Erfe, M.C.; Craft, N.; *et al*. The diversity and host interactions of *Propionibacterium acnes* bacteriophages on human skin. *ISME J.* **2015**, *9*, 2078–2093.
67. Farrar, M.D.; Howson, K.M.; Bojar, R.A.; West, D.; Towler, J.C.; Parry, J.; Pelton, K.; Holland, K.T. Genome sequence and analysis of a *Propionibacterium acnes* bacteriophage. *J. Bacteriol.* **2007**, *189*, 4161–4167.
68. Kuhn, J.; Suissa, M.; Chiswell, D.; Azriel, A.; Berman, B.; Shahar, D.; Reznick, S.; Sharf, R.; Wyse, J.; Bar-On, T.; *et al*. A bacteriophage reagent for Salmonella: Molecular studies on Felix 01. *Int. J. Food Microbiol.* **2002**, *74*, 217–227.
69. Ford, M.E.; Sarkis, G.J.; Belanger, A.E.; Hendrix, R.W.; Hatfull, G.F. Genome structure of mycobacteriophage D29: Implications for phage evolution. *J. Mol. Biol.* **1998**, *279*, 143–164.
70. Schwudke, D.; Ergin, A.; Michael, K.; Volkmar, S.; Appel, B.; Knabner, D.; Konietzny, A.; Strauch, E. Broad-host-range Yersinia phage PY100: Genome sequence, proteome analysis of virions, and DNA packaging strategy. *J. Bacteriol.* **2008**, *190*, 332–342.
71. Garcia, E.; Elliott, J.M.; Ramanculov, E.; Chain, P.S.; Chu, M.C.; Molineux, I.J. The genome sequence of Yersinia pestis bacteriophage φA1122 reveals an intimate history with the coliphage T3 and T7 genomes. *J. Bacteriol.* **2003**, *185*, 5248–5262.
72. Zhao, X.; Cui, Y.; Yan, Y.; Du, Z.; Tan, Y.; Yang, H.; Bi, Y.; Zhang, P.; Zhou, L.; Zhou, D.; *et al*. Outer membrane proteins Ail and OmpF of Yersinia pestis are involved in the adsorption of T7-related bacteriophage Yep-phi. *J. Virol.* **2013**, *87*, 12260–12269.

73. Chang, H.W.; Kim, K.H. Comparative genomic analysis of bacteriophage EP23 infecting *Shigella sonnei* and *Escherichia coli*. *J. Microbiol.* **2011**, *49*, 927–934.
74. De Lappe, N.; Doran, G.; O'Connor, J.; O'Hare, C.; Cormican, M. Characterization of bacteriophages used in the *Salmonella enterica* serovar Enteritidis phage-typing scheme. *J. Med. Microbiol.* **2009**, *58*, 86–93.
75. Hood, A. Phage typing of Staphylococcus aureus. *J. Hyg.* **1953**, *51*, 1–15.
76. Bielke, L.; Higgins, S.; Donoghue, A.; Donoghue, D.; Hargis, B. Salmonella host range of bacteriophages that infect multiple genera. *Poult. Sci.* **2007**, *86*, 2536–2540.
77. Jensen, E.C.; Schrader, H.S.; Rieland, B.; Thompson, T.L.; Lee, K.W.; Nickerson, K.W.; Kokjohn, T.A. Prevalence of broad-host-range lytic bacteriophages of *Sphaerotilus natans*, *Escherichia coli*, and *Pseudomonas aeruginosa*. *Appl. Environ. Microbiol.* **1998**, *64*, 575–580.
78. Olsen, R.H.; Siak, J.S.; Gray, R.H. Characteristics of PRD1, a plasmid-dependent broad host range DNA bacteriophage. *J. Virol.* **1974**, *14*, 689–699.
79. Carson, L.; Gorman, S.P.; Gilmore, B.F. The use of lytic bacteriophages in the prevention and eradication of biofilms of *Proteus mirabilis* and *Escherichia coli*. *FEMS Immunol. Med. Microbiol.* **2010**, *59*, 447–455.

CHAPTER 3

Metagenomics analysis of PYO phage cocktails

In this chapter we present the genomic characterisation and comparison of three batches of the Georgian PYO phage cocktail. The renowned phage researcher, Elizabeth Kutter, from the Evergreen State College, provided us with samples of the PYO cocktail from 1997, 2000, 2010 and 2014 during the Evergreen Phage Meeting of August 2015.

The article has been submitted to the Viruses journal as of 13th of October 2017.

viruses

MDPI

Article

Metagenomic Analysis of Therapeutic PYO Phage Cocktails from 1997 to 2014

Julia Villarroel [1], Mette Voldby Larsen [2], Mogens Kilstrup [3] and Morten Nielsen [1,4,*]

[1] Department of Bio and Health Informatics, Technical University of Denmark, Kemitorvet, Building 208, 2800 Kongens Lyngby, Denmark; julvi@bioinformatics.dtu.dk
[2] GoSeqIt ApS, Ved Klædebo 9, 2970 Hørsholm, Denmark; MVL@goseqit.com
[3] Department of Biotechnology and Biomedicine, Technical University of Denmark, Matematiktorvet, Building 301, 2800 Kongens Lyngby, Denmark; mki@bio.dtu.dk
[4] Instituto de Investigaciones Biotecnológicas, Universidad de San Martín, 1650 San Martín, Buenos Aires, Argentina
* Correspondence: mniel@bioinformatics.dtu.dk; Tel.: +45-4525-2425

Received: 13 October 2017; Accepted: 2 November 2017; Published: 3 November 2017

Abstract: Phage therapy has regained interest in recent years due to the alarming spread of antibiotic resistance. Whilst phage cocktails are commonly sold in pharmacies in countries such as Georgia and Russia, this is not the case in western countries due to western regulatory agencies requiring a thorough characterization of the drug. Here, DNA sequencing of constituent biological entities constitutes a first step. The pyophage (PYO) cocktail is one of the main commercial products of the Georgian Eliava Institute of Bacteriophage, Microbiology and Virology and is used to cure skin infections. Since its first production in the 1930s, the composition of the cocktail has been periodically modified to add phages effective against emerging pathogenic strains. In this paper, we compared the composition of three PYO cocktails from 1997 (PYO97), 2000 (PYO2000) and 2014 (PYO2014). Based on next generation sequencing, de novo assembly and binning of contigs into draft genomes based on tetranucleotide distance, thirty and twenty-nine phage draft genomes were predicted in PYO97 and PYO2014, respectively. Of these, thirteen and fifteen shared high similarity to known phages. Eleven draft genomes were found to be common in the two cocktails. One of these showed no similarity to publicly available phage genomes. Representatives of phages targeting *E. faecalis*, *E. faecium*, *E. coli*, *Proteus*, *P. aeruginosa* and *S. aureus* were found in both cocktails. Finally, we estimated larger overlap of the PYO2000 cocktail to PYO97 compared to PYO2014. Using next generation sequencing and metagenomics analysis, we were able to characterize and compare the content of PYO cocktails separated by 17 years in time. Even though the cocktail composition is upgraded every six months, we found it to remain relatively stable over the years.

Keywords: PYO phage cocktail; metagenomics; human phage therapy

1. Introduction

Phage therapy, the use of phages to cure bacterial infections, has received much attention in recent years due to the emergence and rapid spread of antibiotics resistance. In fact, resistance genes towards last resort treatments for multidrug-resistant bacteria are reported to be circulating all around the world. This highlights an urgent need to coordinate a global effort in the search for antibiotics adjuvants or alternative improved treatments [1–3].

The practice of phage therapy was reported shortly after phage discovery in 1915 [4], when a sudden enthusiasm emerged towards what was believed to be the cure for almost any disease, even before the biology of phages was fully understood [5,6]. The initial excitement rapidly faded,

when phage therapy failed to meet the high expectations, and its practice in western countries soon became obsolete following the discovery of penicillin in 1928 and the advent of the antibiotic era [7].

Despite the displacement of phage therapy by antibiotics in western countries, former Soviet Republics pursued investigations on phages over decades, which today provide a rich trove of knowledge in the field. The related literature has been thoroughly reviewed by Sthephen Abedon et al. [6]. The world leading institution for phage therapy, The Eliava Institute of Bacteriophage, Microbiology and Virology, is located in the former Soviet Republic of Georgia and was founded by the Georgian microbiologist George Eliava in 1923.

In their clinical application, phages are used either as single therapeutic phages, prepared against specific bacterial strains resistant to antibiotics, or phage cocktails which have a broad spectrum of activity towards a set of the most prevalent bacterial strains considered a threat to human health [8]. While the first approach is promoted by the Hirszfeld Institute in Poland, the second is mostly used by the Eliava Institute laboratories, where cocktails' compositions are updated twice per year by adding new phages to target emerging virulent bacteria [5].

Drug regulatory agencies in western countries, European Medicine Agency and Food and Drug Administration (FDA), are expected to require a comprehensive characterization of the components of a cocktail for it to be considered applicable in healthcare. Whole genome sequencing can be deployed for this purpose, along with methods to predict the host of the draft genomes.

In previous studies, the composition of the intestiphage (INTESTI) cocktail from the Georgian Eliava Institute and the ColiProteus cocktail, which is produced by the Russian company Microgen, have been investigated [9,10]. Among other exciting discoveries, these studies identified a new *Proteus* phage genome sequence. However, both studies only examined the composition of a single batch of cocktail and did not look into changes in the composition of the cocktails over years.

In this metagenomic study, we have sequenced and compared the genomic composition of two pyophage (PYO) cocktails, one from 1997, here referred to as PYO97, and the other from 2014, PYO2014. Upon sequencing the DNA of the cocktails and trimming the reads, we assembled the reads into contigs and further binned the contigs from each sample into phage draft genomes. We then compared these draft genomes to phage sequences previously deposited in public databases and examined which draft genomes were common to both samples and in which abundances. Finally, we predicted the host for each phage draft genome. For a third batch of PYO cocktail from 2000 (PYO2000), we were not able to recover phage draft genomes, but we compared the sequence reads to the draft genomes from PYO97 and PYO2014 and found PYO2000 to resemble PYO97 the most.

2. Materials and Methods

Glass vials containing about 10 mL of each of the four PYO phage cocktails—1997, 2000, 2010, and 2014—were kindly provided by Elizabeth Kutter of The Evergreen State College, Olympia, and prepared for sequencing. The bottles are depicted in Figure 1.

Figure 1. The four batches of pyophage (PYO) cocktail investigated here. The glass ampoules are dated (**a**) 1997; (**b**) 2000; (**c**) 2010 and (**d**) 2014.

2.1. DNA Extraction and Library Preparation

The DNA was extracted and isolated using the NORGEN Phage DNA Isolation Kit (Cat. # 46800, Thorold, ON, Canada) following the manual. The extracted DNA was kept at −20 °C until library preparation. PYO97 and PYO2014 had a DNA concentration of 6.06 ng/µL and 1.12 ng/µL, respectively, and a 260/280 ratio within the desired range of 1.8–2.0. PYO2000 and PYO2010 had a DNA concentration of 1.62 ng/µL and 1.61 ng/µL, respectively, but a 260/280 ratio outside the desired 1.8–2.0 range. Due to the 260/280 range, we decided to only process PYO2000 further, and when the resulting sequence reads were of poor quality, refrained from sequencing PYO2010 at all. DNA libraries from PYO97, PYO2000, and PYO2014 were prepared from 1 ng of sample DNA using the NexteraXT Sequencing kit (San Diego, CA, USA) according to the manufacturer's instructions. The resulting libraries were sequenced using the Illumina MiSeq platform (San Diego, CA, USA) yielding 250 bp long paired-end reads.

2.2. Read Trimming

Reads from PYO97, PYO2000, and PYO2014 were checked for quality with Fast Quality Control (FastQC) [11], which produces several different statistics to enable assessment of the quality of short sequence reads. When we, in the following, classified reads as of low or high quality, we based this on the "Per base sequence quality". Reads were trimmed using Prinseq-lite 0.20.4 [12] with the following settings: -trim_qual_right 20 -min_qual_mean 20 -min_len 35 -trim_left 20 -trim_right 10 -derep 14. Non-parallel reads, resulting from trimming, were compensated using cmpfastq [13].

Reads mapping to PhiX174 phage (NC_001422.1), which is used as an internal control in Illumina sequencing, were removed by running MGmapper [14]. MGmapper is a pipeline that takes a fastq file as input and aligns reads to built-in databases using Burrows-Wheeler Alignment algorithm (BWA) [15]. If none of the databases is specified (option -C 0), the program maps the reads to the PhiX174 genome and returns a fastq file of unmapped reads. MGmapper was launched with the following command: MGmapper_PE.pl -i F.fastq -j R.fastq -R -k -C 0 -S.

The reads quality of PYO2000 was low, even after trimming and removal of PhiX174 reads, therefore this sample was excluded from further analyses until we eventually calculated the distances in composition between the three cocktails; see section *Distances in compositions of the cocktails* in Material and Methods.

2.3. Read Mapping

Using Kraken [16], reads from PYO97 and PYO2014 were mapped to the Virus database, which contains complete viral genomes from RefSeq (as of May 2017). Kraken assigns taxonomic labels to metagenomic sequences by searching for exact-matching k-mers (oligonucleotides of length k) between a read and a database of k-mers present in a set of organisms. The Kraken database also stores information about the phylogeny of the organisms. Hence, whenever a query k-mer is present in two or multiple organisms in the database, Kraken assigns the hit to the lowest common ancestor that has these organisms as descendants. Further, reads from PYO97 and PYO2014 were mapped using the Best Mode of MGmapper (option -C) to the built-in databases Bacteria, Archaea, MetaHitAssembly, HumanMicrobiome, Bacteria_draft, Human, Virus, and Fungi downloaded from National Center for Biotechnology Information (NCBI) in June 2017. MGmapper classifies sequences based upon BWA read mapping to a database of reference sequences, allowing for nucleotide variations, inserts and deletions.

2.4. Assembly and Contigs Binning

Reads from PYO97 and PYO2014 were assembled into contigs using the metaSPAdes [17] tool from the SPAdes assembly tool kit (version 3.10.1, Saint Petersburg, Russia) [18] with increasing k-mer lengths (21, 33, 55, 77, 99, 127) as suggested in the software manual.

Metagenome Binning with Abundance and Tetra-nucleotide frequencies (MetaBAT) [19], the software used in this study for binning of contigs into draft genomes, requires the assembly in a fasta file and a sorted Binary Alignment Map (BAM) file as input. Reads from PYO97 and PYO2014 were therefore mapped (BWA 0.7.15) [15] to the respective contigs and the resulting BAM files were sorted using SAMtools sort (SAMtools 1.4) [20]. The assembly fasta file and the sorted BAM file were fed to MetaBAT, v0.32.4 for each sample separately. The samples were binned based on tetranucleotide frequency distance probability. We set the minimum contig length to 2000 bp as previously done [9], the minimum bin size to the minimum that MetaBAT allows, which is 10,000 bp, and the bootstrapping to be run 100 times. MetaBAT was ran in *specific* mode: –p1 90 –p2 90 –pB 30 –minProb 80 –minBinned 40 –minCorr 96, to minimize contigs belonging to different phages being binned together.

2.5. Finding the Most Similar Reference Genome

Phage whole genome sequences (WGS) were downloaded from the NCBI viral RefSeq database [21] and PhAnToMe [22] resulting in 3889 unique WGS as of May 2017. To find the closest

reference to each bin from PYO97 and PYO2014, we ran MetaPhinder [23]. MetaPhinder is a Blast-based method, which for a given query entry provides a measure, the percentage Average Nucleotide Identity (ANI), that integrates multiple hits of the query genome to all sequences in a database. The ANI value is calculated as

$$\%ANI = \frac{\sum_{i=1}^{n} id_i * al_i}{\sum_{i=1}^{n} al_i} * m_{cov} \quad (1)$$

where n is the number of Blastn hits between the query sequence and all sequences in the database with an e-value of 0.05 or smaller, id is the Blastn % identity value between the query and a given database hit, al is the corresponding Blastn alignment length, and m_{cov} is the coverage of the query sequence over all hits. Using this approach, a Blast database was constructed from each bin and next queried with each of the 3889 phage Whole Genome Sequences (WGS). For each bin-database, MetaPhinder reported the ANI for each query WGS, and the query with the highest ANI was selected as the one matching the bin the most.

2.6. Checking Consistency within and between Bins

The trimmed reads devoid of PhiX174 of PYO97 and PYO2014 were aligned to the respective contigs using BWA [15]. The coverage, here the number of reads mapping to the contigs times the read length divided by the length of the contigs in bp, was calculated using samtool depth [20]. If high variance of coverage values were observed for the member contigs of a particular bin, the bin was manually split into smaller bins, each only containing contigs with a confined range of coverage values.

Bins that shared the best matching genome among the 3889 WGS had a similar coverage and no overlapping contigs between them were manually merged.

2.7. Bin Annotation

To classify if a given bin is a phage or not, we estimated the ANI of each bin from PYO97 and PYO2014 towards the Blastn database of the 3889 phage WGS described earlier. An ANI threshold of 10% was chosen to discriminate between phage and non-phage query bins. For bins containing more than one contig, a weighted ANI average was calculated as

$$\overline{ANI} = \frac{\sum_{i=1}^{n} ANI_i * l_i}{\sum_{i=1}^{n} l_i} \quad (2)$$

where n is the number of contigs in the bin and l is the length of the member contigs. HostPhinder [24] was used to predict the bacterial host of the draft genomes. HostPhinder predicts the host of a phage genome sequence by searching for overlapping 16-mers between the query and a database of phage genomes with an annotated host. Upon finding the best matching hits in the database, HostPhinder predicts the host to be the most represented host among the top hits. The prediction is associated with a reliability score from 0 to 1. Only scores higher than 0.1 are considered reliable [24]; we therefore only reported results above this threshold.

2.8. Similarities between PYO97 and PYO2014.

To estimate the similarity between bins of PYO97 and PYO2014, MetaPhinder was used as follows: A Blast database of contigs from a given bin from one sample was searched with each contig of a bin of the other sample (the query bin). Next, the query bin was assigned a weighted mean ANI calculated from the ANIs and lengths (l) of query contigs, Equation (2). For each database bin, the query bin with the highest $\overline{ANI}$ was considered the matching candidate. The reciprocal ANI was calculated using OrthoANI [25], which takes into account only orthologous fragment pairs between the two sequences.

2.9. Phage Draft Genome Visualization

Phage draft genomes were visualized using BLAST Ring Image Generator v0.95 (BRIG) [26]. Alternatively, we ran a customized python script to produce xml files from Blast results and used CGView Java Package to visualize them as circular genomes [27].

2.10. Bin Classification

Bins were classified into six categories according to high (ANI $\geq$ 70%) or medium/low (ANI < 70%) resemblance to a reference genome or to a bin in the other sample. Bins that were more than 10% longer than the best matching reference genome and that included overlapping contigs were classified as *collapsed bins*. Bins with ANI < 10% towards phages in public databases were labeled as *special cases*. Bins composed by more than 20 contigs which were shorter than 7000 bp were too fragmented to be considered drafts of genomes and were therefore also designated as special cases. For simplicity, here we will refer to draft genomes to indicate bins that are not special cases. When the term *bin* is used, then all bins including special cases are intended.

2.11. Phage Abundances

To further check whether the phages of one cocktail sample were present in the other and with which relative abundance, we mapped the PYO97 reads to the PYO2014 bins and vice versa using BWA. The bin coverage values, calculated here as the number of reads mapping to the bins times the read length and divided by the length of the bins in bp, were obtained using samtools depth [20].

2.12. Distances in Compositions of the Cocktails

We ran Mash v1.1.1 [28] to determine the distances in terms of composition between the samples. Trimmed reads devoid of PhiX174 of samples PYO97, PYO2000 and PYO2014 were used.

Mash enables the comparison of metagenomic samples by splitting them into constituent k-mers and reducing the samples into sketches of representative k-mers. From these size-reduced sketches, Mash can rapidly calculate the Jaccard index based on co-occurring k-mers. Based on the Jaccard index, Mash estimates global mutation distances ($0 \leq D \leq 1$) between samples. The results have a strong correlation with the ANI. We chose a k-mer size of 16, a sketch size of 400 and a minimum of 2 copies of k-mers in order for the k-mer to be considered as a candidate for the sample sketch. Mash was launched as follows:

```
mash sketch -m 2 -k 16 -s 400 -o distance.msh tmp/*.fq

mash dist distance.msh distance.msh > distances.tab
```

where tmp/*.fq represents the folder containing the fastq files of interleaved reads for the 3 samples.

To get the bootstrap mean and confidence interval of the distances, pair reads of the 3 samples were separately shuffled with resampling 100 times. In each resampling, Mash made sketches of the 3 samples and calculated pairwise distances between the samples. This resulted in one hundred 3×3 distance tables from which the mean and mean squared error of each pairwise distance were calculated.

3. Results

3.1. Reads Statistics

The DNA from each of the four batches of PYO cocktail was extracted. The yield from PYO2010 was very low and we, accordingly, chose not to sequence it. Table 1 reports the number of reads before and after trimming and removal of PhiX174 reads obtained from PYO97, PYO2000, and PYO2014.

Table 1. Reads statistics.

Sample	# Reads	# Reads after Trimming	# Reads after Removing PhiX174
PYO97	5,228,884 (1,280,000 kbp)	2,035,496 (420,000 kbp)	1,965,233 (410,000 kbp)
PYO2000	1,648,430 (410,000 kbp)	1,366,749 (300,000 kbp)	1,110,522 (240,000 kbp)
PYO2014	18,240,556 (4,470,000 kbp)	6,660,081 (1,380,000 kbp)	6,577,613 (1,370,000 kbp)

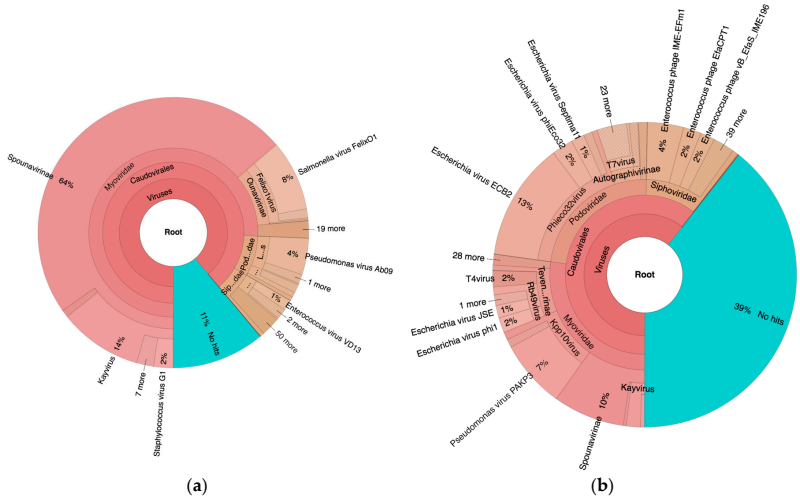
Means "number of".

PYO2000 was shown to have poor read quality, with a per base sequence quality significantly lower than PYO97 and PYO2014. On account of this, we only attempted to generate phage draft genomes for PYO97, the first time point and PYO2014, the last time point. The trimmed reads devoid of PhiX174 of PYO2000 were mapped to the draft genomes of PYO97 and PYO2014 to examine genomic overlap; see Material and Methods and the section *Phage abundance and bin comparison* in the Results.

3.2. Reads Mapping

To get an overview of what was present in the PYO97 and PYO2014 cocktails, reads were initially mapped to the Kraken Virus database.

As seen in Figure 2, 89% and 61% of the reads mapped to viruses of the order *Caudovirales* in PYO97 and PYO2014, respectively. Of these, most mapped to the family *Myoviridae* (85%), while 9% and 6% mapped to *Podoviridae* and *Siphoviridae*, respectively, for PYO97. The ratios of represented phage families within the order *Caudovirales* in PYO2014 were more even: 45% *Myoviridae*, 38% *Podoviridae*, and 17% *Siphoviridae*.



(**a**) (**b**)

Figure 2. Krona map based on reads from PYO97 (**a**) and PYO2014 (**b**) mapped to the Kraken Virus database. Interactive charts can be found at https://julvi.github.io/PYO97_krona.html and https://julvi.github.io/PYO2014_krona.html for the respective samples.

For PYO97, 11% of the reads did not map to any of the sequences in the Kraken Virus database. The corresponding number for PYO2014 was 39%.

To examine if the unmapped reads from above mapped to sequences from other organisms than viruses, MGmapper was next ran using the databases of Bacteria, Archea, MetaHitAssembly [29],

HumanMicrobiome [30], Bacteria draft, Human, Viruses and Fungi. No significant mapping to other databases besides Viruses was reported, Table S1.

3.3. Assembly and Contigs Binning

In order to detect any draft genome that was common between PYO97 and PYO2014, we proceeded in the assembly and downstream analysis of the two samples with high quality reads.

The assembly yielded 179 and 270 contigs longer than 2000 bp for PYO97 and PYO2014, respectively (Table 2). Note, that while the 270 contigs from PYO2014 in total encompass 2759 kbp to which 6,516,794 reads map, the 179 contigs from PYO97 encompass 3034 kbp to which only 1,924,746 reads map, indicating that the depth of coverage obtained for the PYO97 cocktail is not as high as for the PYO2014 cocktail.

Table 2. Summary of the assembly results.

Sample	# Contigs	Longest and Shortest Contig	# Contigs Longer than 2 kbp–Percentage of Reads Mapping to the Contigs
PYO97	3005 (4403 kbp)	169,029 bp (128 bp)	179 (3034 kbp)–97.9%
PYO2014	4165 (4865 kbp)	282,352 bp (128 bp)	270 (2759 kbp)–99.1%

Means "number of".

The assembly of metagenome reads often fails to produce entire genomes even for small phage genomes. To arrive at a more complete assembly, MetaBAT was used to group contigs with similar tetranucleotide frequency, allowing to come close to what can be considered draft genomes. MetaBAT produced 33 bins from PYO97 and 31 from PYO2014 and were able to bin more than 90% bp of the contigs longer than 2000 bp for each sample (Table 3).

Table 3. Number of bins yielded by MetaBAT and number and percentage of binned bp out of the total number of bp in contigs larger than 2000 bp.

Sample	# Bins	# Binned bp (Percentage of Binned bp)
PYO97	33	2,735,811 (90.16%)
PYO2014	31	2,494,104 (90.39%)

Means "number of".

3.4. Consistency within and between Bins

The bins produced by MetaBAT were composed of between 1 and 50 contigs. In cases where some of the binned contigs overlapped when mapped to the reference sequence, an effort was made to split the bin according to the differences in contig coverage values. After such splits, the newly formed bin generally had a different closest reference genome to the original bin. An illustration of this is shown in Figure 3a. Here, the bin PYO97_10, with *Escherichia* phage PBECO 4 as the closest reference, was split into PYO97_10_85.139.47.48.59.38.5.35.15.78.55.44.14 with the same reference as the original bin and PYO_10_3.8.10.28.42, which in turn had *Escherichia* phage 121Q as the closest reference genome.

Bins mapping to the same reference were merged, if their coverage was in the same range. An example of this is shown in Figure 3b. Here, three bins from PYO97, PYO97_22, PYO97_3 and PYO97_14, which shared a high sequence similarity to *Salmonella* phage Shivani and had coverage values between 58 and 72, were merged into a single bin PYO97_22.3.14 which preserved the reference genome and showed a coverage of 65 with a lower mean standard deviation compared to the original bins. This and other examples of bin merging are listed in Table 4. PYO2014_3.16.29, in our view, represents two or more closely related phages (see Figure 3c), that are identical in the region represented by PYO2014_29, but slightly differ in the regions represented by PYO2014_3 and PYO2014_16.

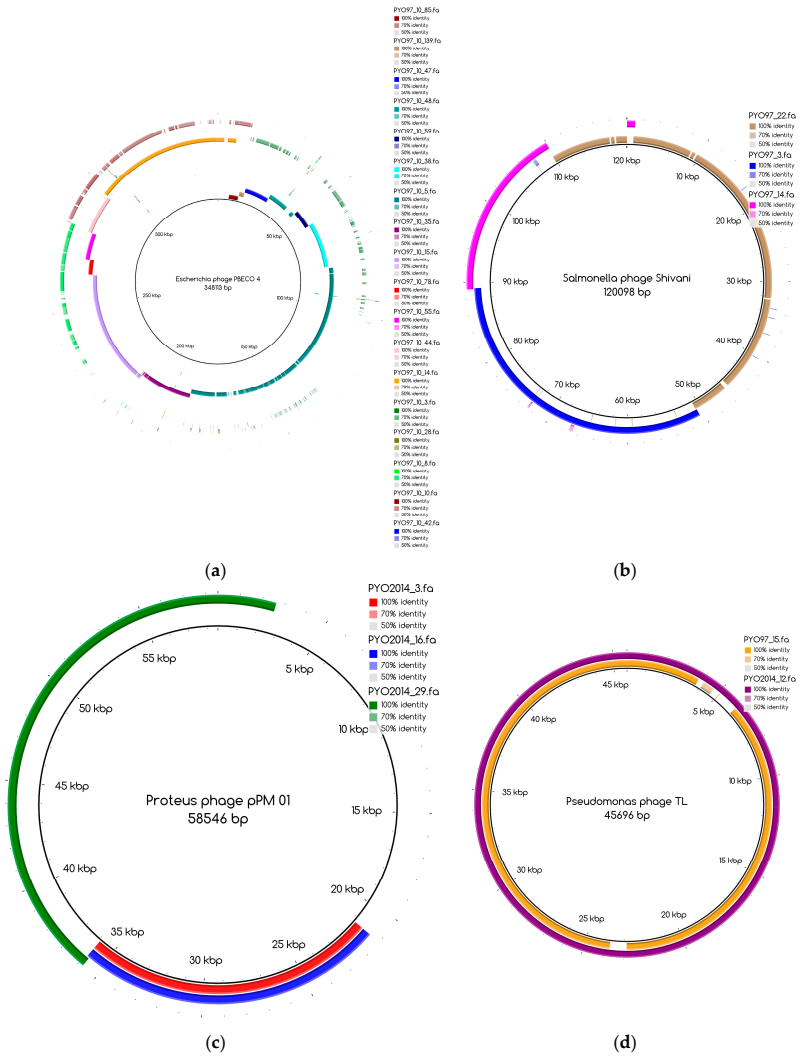


Figure 3. *Cont.*

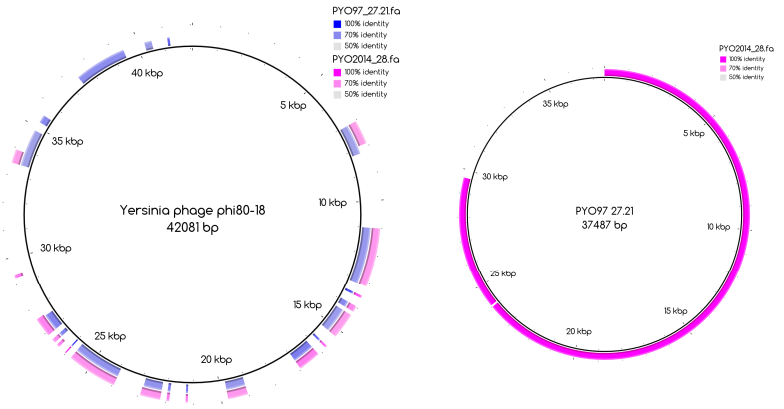


Figure 3. Blast Atlases. (**a**) Example of an original bin, PYO97_10 (coverage 13.9 ± 0.015), that was split into two bins: one made of the contiguous contigs closest to the reference in the middle until the orange contig, bin PYO97_10_85.139.47.48.59.38.5.35.15.78.55.44.14 (coverage 12.4 ± 0.014), and the second one, bin PYO97_10_3.8.10.28.42 (coverage 16.9 ± 0.025), containing the remaining contigs; (**b**) Merging of three bins PYO97_22, PYO97_3 and PYO97_14 into one PYO97_22.3.14 which covers the entire reference genome; (**c**) Collapsed bin in PYO2014; these three contigs have been grouped together to form a collapsed bin. The difference between a normal bin and a collapsed bin is the presence of overlapping contigs in the latter probably derived from shared sequences between species of the same phage family; (**d**) Corresponding draft genomes from the two samples aligning to the reference, *Pseudomonas* phage TL; (**e**) PYO97_27.21 and PYO2014_28 are highly similar and do not resemble any known sequence; (**f**) Alignment of PYO2014_28 to PYO97_27.21.

Table 4. The bins to be merged are indicated in the first three columns. The fourth and fifth columns show the resulting merged bin and the closest reference, respectively. Coverage values are in parentheses.

Bin to be Merged 1	Bin to be Merged 2	Bin to be Merged 3	Merged Bin	Shared Reference
PYO97_17 (44.67 ± 0.05)	PYO97_30 (53.82 ± 0.07)		PYO97_17.30 (46.65 ± 0.04)	*Salmonella* phage SSE-121 (NC_027351.1)
PYO97_3 (57.92 ± 0.08)	PYO97_14 (66.42 ± 0.14)	PYO97_22 (71.55 ± 0.14)	PYO97_22.3.14 (65.41 ± 0.08)	*Salmonella* phage Shivani (NC_028754.1)
PYO2014_3 (2294.3 ± 3.8)	PYO2014_16 (2294.98 ± 3.81)	PYO2014_29 (2222.82 ± 2.54)	PYO2014_3.16.29 (1659.65 ± 2.65)	*Proteus* phage pPM_01 (NC_028812.1)
PYO2014_13 (109.82 ± 0.21)	PYO2014_26 (120.43 ± 0.16)		PYO2014_26.13 (118.21 ± 0.14)	*Pseudomonas* phage PEV2 (NC_031063.1)

Eventually, after this manual splitting and merging of bins, 30 and 29 final bins were obtained from PYO97 and PYO2014, respectively.

Two bins of PYO2014 were both composed of 26 contigs, all shorter than 7000 bp. Due to this fragmentation, they were labeled as special cases.

3.5. Bin Annotation

We calculated the ANI of the final bins towards the set of publicly available phage genomes to discriminate between phage bins (bins similar to previously sequenced phages) and non-phage bins (bins that share little similarity to known phage sequences). We chose a very stringent threshold of 10% ANI to classify a bin as of phage origin. Using this threshold, five (17%) and three (10%) bins from PYO97 and PYO2014, respectively, were classified as non-phages and added to the special cases. One of the non-phage bins from PYO2014 was already a special case due to its fragmentation, see above. Bins not belonging to the special cases will hereafter be referred to as draft genomes.

To further characterize the draft genomes, we predicted their bacterial hosts using HostPhinder. Table 5 reports the predicted represented bacterial hosts in the two samples.

Table 5. Number of phage draft genomes from the two samples which HostPhinder predicted to infect the respective host. Only results having a score higher than the reliability threshold of 0.1 are reported.

Bacterial Host	# Representative Phage Draft Genomes in PYO97	# Representative Phage Draft Genomes in PYO2014
Enterococcus faecalis	2	3
Enterococcus faecium	1	1
Escherichia coli	4	7
Klebsiella pneumoniae	1	0
Proteus mirabilis	2	1
Pseudomonas aeruginosa	3	4
Salmonella enterica	2	0
Salmonella enteritidis	2	0
Shigella sonnei	0	1
Staphylococcus aureus	1	1

Means "number of".

3.6. Similarities between PYO97 and PYO2014

Approximately every 6 months, the Eliava Institute laboratories update the content of the PYO cocktail to cope with the emergence of new clinically problematic bacterial strains. New effective phages are added, while phages added in previous batches slowly dilute, leading to an overall change of the cocktail composition.

We investigated how much overlap in the compositions of PYO97 and PYO2014 was appreciable by looking for common phage draft genomes between the two cocktails. The corresponding pairs of draft genomes between the two samples were determined using MetaPhinder in a pairwise manner as described in Materials and Methods.

Table 6 reports the pairs identified by MetaPhinder, where at least one of the ANI, calculated either by using PYO97's or PYO2014's phage drafts as databases was higher than 70%.

Table 6. Overview of correspondent draft genomes between PYO97 and PYO2014 and the reciprocal ANI. The last column displays the targeted host as predicted by HostPhinder. Bins 10_85…, and 27_42… in the table correspond to bin 10_85.139.47.48.59.38.5.35.15.78.55.44.14 and 27_42.21.133.90.116.20.14.109.73, respectively.

Bin from PYO97	Bin from PYO2014	Reciprocal ANI (%)	Predicted Targeted Bacterial Host
16	30	99.9	*S. aureus*
27.21	28	98.6	*Yersinia enterocolitica* *
11	26.13	97.2	*P. aeruginosa*
12	1	98.8	*E. faecium*
25	3.16.29	99.5	*Proteus mirabilis*
15	12	98.7	*P. aeruginosa*
29	21	96.2	*E. coli*/*Shigella sonnei* **

Table 6. *Cont.*

Bin from PYO97	Bin from PYO2014	Reciprocal ANI (%)	Predicted Targeted Bacterial Host
4	27_42…	98.1	*E. faecalis*
8	25	88.4	*E. coli*
10_85…	31	89.6	*E. coli*
23	20	85.4	*E. faecalis*

* Indicates that the prediction by HostPhinder had a low score and was hence unreliable. ** In this case, HostPhinder predicted a different host for each draft genome.

The combined results of HostPhinder and pairwise MetaPhinder displayed in Table 6 strongly suggest that the same phages against *E. faecalis* (2), *E. faecium* (1), *E.coli* (2), *P. mirabilis* (1), *P. aeruginosa* (2), and *S. aureus* (1) are present in both samples; where the numbers in parenthesis are the counts of likely identical phages found in both samples which are capable of infecting the specified host.

3.7. Draft Genomes Classification

According to their similarity to reference genomes and to the presence of a likely counterpart at the other time point (see Materials and Methods), draft genomes were classified within the categories listed in Table 7. The special cases include highly fragmented bins and non-phage bins. For these reasons, special cases are referred to as bins and not as draft genomes. Table 7 also displays the number of draft genomes/bins from each sample belonging to each category. As an illustrative example, the six draft genomes from PYO97 in category 1, have high similarity to a reference genome and to draft genomes in PYO2014. One example of pairs of corresponding draft genomes is given by PYO97_15 and PYO2014_12, Figure 3d. The two draft genomes share high similarity to Pseudomonas phage TL.

Table 7. Count of draft genomes/bins belonging to each category.

Class	PYO97	PYO2014
(1) Near-complete draft genome with high resemblance to reference phage and counterpart in the other sample.	6	4
(2) Near-complete draft genome with high resemblance to reference phage, but no counterpart in the other sample.	5	8
(3) Partial draft genome with low/medium resemblance to reference phage and counterpart in the other sample.	1	1
(4) Partial draft genome with no resemblance to reference phage and no counterpart in the other sample.	11	8
(5) Collapsed bins.	2	4
(6) Special cases, including highly fragmented bins and bins classified as non-phages.	5	4

The number of draft genomes belonging to each category does not necessarily match between the two samples, even for the categories of draft genomes with a counterpart in the other sample, categories 1 and 3. This is, for instance, the case for draft genome PYO97_29, category 1, mapping to the collapsed draft genome PYO2014_21, which belongs to the fifth category, Figure A1. Tables 8 and 9 provide a general overview of the phage draft genomes found in PYO97 and PYO2014, respectively, together with an indication of the most likely taxonomic group they belong to. For a more thorough description of the draft genomes in each category, see Tables S2 and S3 for PYO97 and PYO2014, respectively. A case worth noticing is that of the draft genomes PYO97_27.21 and PYO2014_28 in category 3. These draft genomes share similarity with ANI > 70%, but have low ANI to the common reference genome, *Yersinia* phage phi80-18 (refer to, Figure 3e,f for an illustration of the overlap between the two bins). This could suggest that the PYO97_27.21 and PYO2014_28 draft genomes represent a previously uncharacterized phage.

Table 8. Overview of the phage draft genomes and bins of PYO97 indicating the most likely taxonomic group they belong to. PYO97_10_85 … and PYO97_10_3… correspond to PYO97_10_85.139.47.48.59.38.5.35.15.78.55.44.14 and PYO97_10_3.8.10.28.42, respectively.

Bin Name	# Contigs	Size (bp)	Closest Relative in the Database	ANI (%)	Most Likely Taxonomic Group
PYO97					
PYO97 near-complete draft genomes with high resemblance to reference phage and counterpart in PYO2014. (Category 1)					
PYO97_4	1	149,561	*Enterococcus* phage EFDG1 (NC_029009.1)	89.77	*Caudovirales; Myoviridae; unclassified Myoviridae*
PYO97_10_85…	13	344,749	*Escherichia* phage PBECO 4 (NC_027364.1)	90.848	*Caudovirales; Myoviridae; unclassified Myoviridae*
PYO97_11	1	72,136	*Pseudomonas* phage PEV2 (NC_031063.1)	97.37	*Caudovirales; Podoviridae; N4likevirus; unclassified N4likevirus*
PYO97_15	1	44,667	*Pseudomonas* phage TL (NC_023583.1)	92.02	*Caudovirales; Podoviridae; Luz24virus; Pseudomonas virus TL*
PYO97_16	1	130,932	*Staphylococcus* phage Sb-1 (HQ163896.1)	96.86	*Caudovirales; Myoviridae; Spounavirinae; Spo1virus; unclassified SPO1-like viruses*
PYO97_29	1	169,029	*Shigella* phage SHFML-11 (NC_030953.1)	89.959	*Caudovirales; Myoviridae; Tevenvirinae; T4virus; unclassified T4virus*
PYO97 near-complete draft genomes with high resemblance to reference phage, but no counterpart in PYO2014. (Category 2)					
PYO97_7	7	166,126	*Klebsiella* phage vB KpnM KpV477 (NC_031087.1)	88.66	*Caudovirales; Myoviridae*
PYO97_8	1	38,419	*Enterobacteria* phage 285P (NC_015249.1)	79.568	*Caudovirales; Podoviridae; Autographivirinae; T7virus; unclassified T7-like viruses*
PYO97_22.3.14	3	109,428	*Salmonella* phage Shivani (NC_028754.1)	95.33	*Caudovirales; Siphoviridae; T5virus; Salmonella virus Shivani*
PYO97_24	1	44,541	*Proteus* phage PM 85 (NC_027379.1)	92.726	*Caudovirales; Podoviridae; unclassified Podoviridae*
PYO97_32	3	47,235	*Salmonella* phage vB SenS-Ent1 (HE775250.1)	86.967	unclassified
PYO97 partial draft genome with low/medium resemblance to reference phage and counterpart in PYO2014. (Category 3)					
PYO97_27.21	2	37,487	*Yersinia* phage phi80-18 (NC_019911.1)	22.104	*Caudovirales; Podoviridae*
PYO97 partial draft genomes with no resemblance to reference phage and no counterpart in PYO2014. (Category 4)					
PYO97_1	1	11,445	*Escherichia* phage vB EcoM AYO145A (NC_028825.1)	10.99	*Caudovirales; Myoviridae*
PYO97_5	3	29,155	*Pseudomonas* phage vB Pae-TbilisiM32 (JQ307386.1)	68.72	*Caudovirales; Podoviridae; Autographivirinae*
PYO97_9	1	10,727	*Salmonella* phage BP63 (NC_031250.1)	19.779	*Caudovirales; unclassified Caudovirales*
PYO97_10_3…	5	343,801	*Escherichia* phage 121Q (NC_025447.1)	28.408	*Caudovirales; Myoviridae*
PYO97_13	1	37,843	*Hamiltonella* virus APSE1 (NC_000935.1)	9.777	*Caudovirales; Podoviridae*
PYO97_17.30	7	90,209	*Salmonella* phage SSE121 (NC_027351.1)	58.832	*Caudovirales; Myoviridae; Vequintavirinae*
PYO97_20	1	90,712	*Cronobacter* phage vB CsaP GAP52 (NC_019402.1)	19.54	*Caudovirales; Podoviridae*
PYO97_25	1	25,293	*Proteus* phage pPM_01 (NC_028812.1)	41.01	*Caudovirales; Siphoviridae; unclassified Siphoviridae*
PYO97_26	5	171,908	Cronobacter phage S13 (NC_028773.1)	45.28	*Caudovirales; Myoviridae; unclassified Myoviridae*
PYO97_28	5	30,952	*Salmonella* phage 21 (NC_029050.1)	21.43	*Caudovirales; Myoviridae*
PYO97_31 *	3	69,885	*Salmonella* phage *Felix* 01 (NC_005282.1)	75.359	*Caudovirales; Myoviridae; Ounavirinae*

Table 8. *Cont.*

Bin Name	# Contigs	Size (bp)	Closest Relative in the Database	ANI (%)	Most Likely Taxonomic Group
PYO97 collapsed bins. (Category 5)					
PYO97_12	5	55,452	*Enterococcus* phage IME-EFm5 (NC_028826.1)	69.288	*Caudovirales; Siphoviridae; unclassified Siphoviridae*
PYO97_23	5	73,434	*Enterococcus* phage VD13 (NC_024212.1)	74.273	*Caudovirales; Siphoviridae; Sap6virus*
PYO97 special cases, including bins classified as non-phages. (Category 6)					
PYO97_2	1	11,313	uncultured Mediterranean phage uvMED-GF-C25 -MedDCM-OCT-S33-C258 (AP014078.1)	0.704	unknown
PYO97_6	8	23,397	uncultured Mediterranean phage uvMED-CGF-C14B -MedDCM-OCT-S36-C258 (AP013800.1)	1.426	unknown
PYO97_18	1	11,354	*Pseudomonas* phage PRR1 (NC_008294.1)	0.984	unknown
PYO97_19	10	284,533	*Staphylococcus* phage Sb-1 (HQ163896.1)	18.88	*Caudovirales; Myoviridae*
PYO97_33	3	10,088	uncultured Mediterranean phage uvMED-CGF-C23 -MedDCM-OCT-S24-C232 (AP013582.1)	1.131	unknown

* PYO97_31 is 20 kbp shorter than the reference, therefore it was placed in this category, despite the high ANI of the reference genome. # Means "number of". The 5th column reports the ANI of the reference genome towards the bin.

Table 9. Overview of the phage draft genomes and bins of the cocktail PYO2014, indicating the most likely taxonomic group they belong to. PYO2014_27_42… corresponds to PYO2014_27_42.21.133.90.116.20.14.109.73.

Bin Name	# Contigs	Size (bp)	Closest Relative in the Database	ANI (%)	Most Likely Taxonomic Group
PYO2014					
PYO2014 near-complete draft genomes with high resemblance to reference phage and counterpart in PYO97. (Category 1)					
PYO2014_1	1	42,721	*Enterococcus* phage IME-EFm5 (NC_028826.1)	70.16	*Caudovirales; Siphoviridae; unclassified Siphoviridae*
PYO2014_12	1	47,209	*Pseudomonas* phage TL (NC_023583.1)	97.91	*Caudovirales; Podoviridae; Luz24virus; Pseudomonas virus TL*
PYO2014_27_42…	9	138,228	*Enterococcus* phage EFDG1 (NC_029009.1)	81.548	*Caudovirales; Myoviridae; unclassified Myoviridae*
PYO2014_30	1	138,269	*Staphylococcus* phage ISP (FR852584.1)	99.36	*Caudovirales; Myoviridae; Spounavirinae; Kayvirus; Staphylococcus virus G1*
PYO2014 near-complete draft genomes with high resemblance to reference phage, but no counterpart in PYO97. (Category 2)					
PYO2014_2	1	76,529	*Escherichia* phage ECBP2 (NC_018859.1)	77.91	*Caudovirales; Podoviridae; Phieco32virus; Escherichia virus ECB2*
PYO2014_4	1	282,352	*Pseudomonas* phage phiKZ. (NC_004629.1)	94.53	*Caudovirales; Myoviridae; Phikzvirus*
PYO2014_8	1	36,807	*Enterococcus* phage EFAP-1 (NC_012419.1)	74.45	*Caudovirales; Siphoviridae; unclassified Siphoviridae*
PYO2014_17	1	88,099	*Pseudomonas* phage CHA P1 (NC_022974.1)	94.91	*Caudovirales; Myoviridae*
PYO2014_18	1	147,760	*Enterobacteria* phage phi92 (NC_023693.1)	91.5	*Caudovirales; Myoviridae; unclassified Myoviridae*
PYO2014_23	1	38,847	*Enterobacteria* phage K1F (NC_007456.1)	82.27	*Caudovirales; Podoviridae; Autographivirinae; T7virus; unclassified T7-like viruses*
PYO2014_26.13	2	65,818	*Pseudomonas* phage PEV2 (NC_031063.1)	90.705	*Caudovirales; Podoviridae; Lit1virus; Pseudomonas virus Ab09*
PYO2014_27_16.5	2	139,828	*Enterococcus* phage EFLK1 (NC_029026.1)	90.812	*Caudovirales; Myoviridae; unclassified Myoviridae*

Table 9. *Cont.*

Bin Name	# Contigs	Size (bp)	Closest Relative in the Database	ANI (%)	Most Likely Taxonomic Group
PYO2014 partial draft genome with low/medium resemblance to reference phage and counterpart in PYO97. (Category 3)					
PYO2014_28	1	33,115	*Yersinia* phage phi80-18 (NC_019911.1)	16.79	*Caudovirales; Podoviridae; Autographivirinae*
PYO2014 partial draft genomes with no resemblance to reference phage and no counterpart in PYO97. (Category 4)					
PYO2014_7	1	103,078	*Escherichia* phage bV EcoS AKFV33 (HQ665011.1)	14.32	*Caudovirales; Siphoviridae*
PYO2014_9	1	37,468	*Enterococcus* phage vB IME197 (NC_028671.1)	15.18	*Caudovirales; Siphoviridae*
PYO2014_19	1	43,272	*Pseudomonas* phage vB PaeP Tr60 Ab31 (NC_023575.1)	45.35	*unclassified dsDNA phage*
PYO2014_10	1	10,736	*Escherichia* phage PBECO 4 (NC_027364.1)	3.06	*Caudovirales; Myoviridae*
PYO2014_11	1	13,190	*Escherichia* phage PE3-1 (NC_024379.1)	29.52	*Caudovirales; Podoviridae; Autographivirinae*
PYO2014_14	1	17,615	*Escherichia* phage PBECO 4 (NC_027364.1)	4.35	*Caudovirales; Myoviridae*
PYO2014_20	4	16,677	*Enterococcus* phage VD13 (NC_024212.1)	20.55	*Caudovirales; Siphoviridae*
PYO2014_31	50	227,129	*Escherichia* phage PBECO 4 (NC_027364.1)	54.417	*Caudovirales; Myoviridae*
PYO2014 collapsed bins. (Category 5)					
PYO2014_3.16.29	3	54,712	*Proteus* phage pPM_01 (NC_028812.1)	64.733	*Caudovirales; Siphoviridae; unclassified Siphoviridae*
PYO2014_5	25	193,706	*Enterobacteria* phage GEC-3S (NC_025425.1)	90.11	*Caudovirales; Myoviridae; Tevenvirinae; T4virus*
PYO2014_21	22	180,343	*Shigella* phage SHFML-11 (NC_030953.1)	88.2	*Caudovirales; Myoviridae; Tevenvirinae; T4virus*
PYO2014_25	3	78,290	*Enterobacteria* phage 285P (NC_015249.1)	76.539	*Caudovirales; Podoviridae; Autographivirinae; T7virus*
PYO2014 special cases, including bins classified as non-phages. (Category 6)					
PYO2014_6	26	75,778	Uncultured phage WW-nAnB strain 2 (NC_026612.1)	1.91	unknown
PYO2014_15	1	20,152	uncultured Mediterranean phage uvMED-CGF-C24-MedDCM-OCT-S28-C185 (AP013656.1)	0.69	unknown
PYO2014_22	5	57,117	*Pseudomonas* phage O4 (NC_031274.1)	1.78	*dsDNA viruses, no RNA stage*
PYO2014_24	26	89,259	*Cronobacter* phage vB CsaM GAP161 (NC_019398.1)	42.11	*Caudovirales; Myoviridae; Tevenvirinae*

Means "number of". The 5th column reports the ANI of the reference genome towards the bin.

It is worth noticing that the percentage of reads that align to the bins with ANI < 40 towards known sequences was 6.87% and 22.79% for PYO97 and PYO2014, respectively. These percentages align with the differences in percentages of unclassified reads between the two samples, as found when using Kraken in paragraph 3.2: 11% for PYO97 and 39% for PYO2014. However, the results from BWA and Kraken analyses are not directly comparable since BWA alignment allows for indels and point mutation [15], while Kraken only reports exact matching k-mers [16].

3.8. Phage Abundances and Bin Comparison

To estimate the relative abundances of bins in PYO97 and PYO2014, we calculated the bin coverage of the PYO97's and PYO2014's bins by the reads of the samples PYO97 and PYO2014. To account for the difference in the number of reads between sample PYO97 and PYO2014, we normalized the coverage values by the total number of reads of the respective sample.

The distribution of the bins according to the bin coverage by the reads of PYO97 and PYO2014 is shown in Figure 4. Circles represent draft genomes listed in Table 6 having a counterpart in the other

sample. These draft genomes had generally high abundances in both samples, which is deducible from the position of circle data points in the top right corner of the graph. PYO97_27.21 and PYO2014_28 offer an interesting example, as these two draft genomes are almost completely overlapping in terms of relative abundance in the two samples. As stated earlier, these two draft genomes have high ANI and both had low similarity to the common best reference, *Yersinia* phage phi80-18. HostPhinder predicted *Yersinia enterocolitica* to be the host of both, yet with a low confidence, see last column in Tables S1 and S2. Figure 3f displays the sequence similarity between the two bins. These results thus further support the conclusion that this phage draft is an example of a previously unsequenced phage genome.

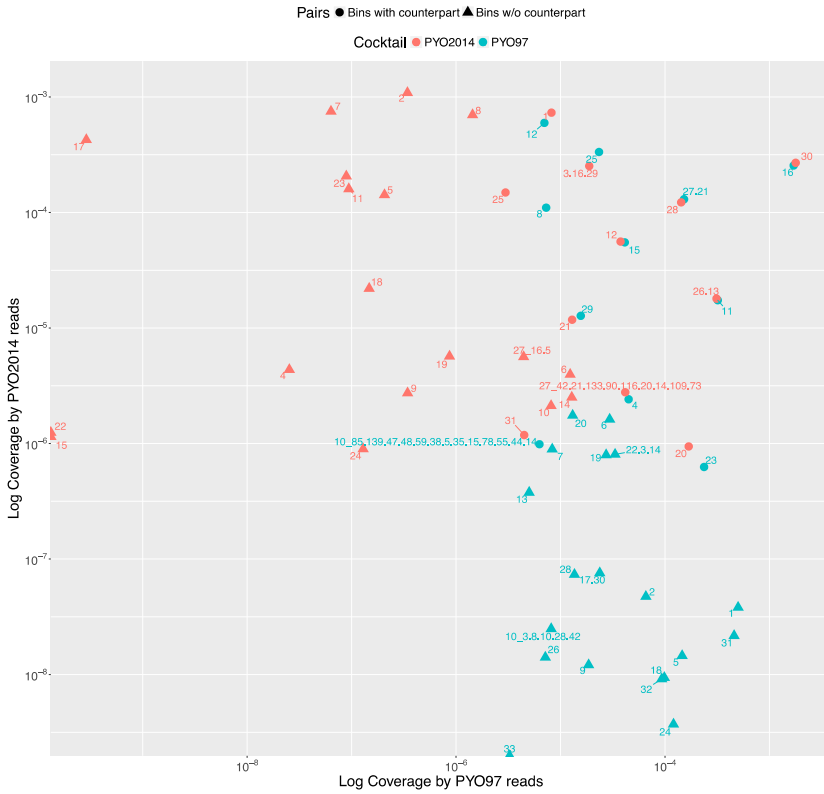


Figure 4. All the bins from PYO97 and PYO2014 are plotted according to the Log to base 10 of the coverage by PYO97's reads, x-axis and PYO2014's reads, y-axis. Bins from PYO97 are depicted in blue, whilst bins from PYO2014 are red. Circle shaped data points represent pairing bins between the two samples, i.e., bins for which MetaPhinder found a matching counterpart in the other sample with a ANI > 70%; see Table 6 and Material and Methods, *Estimate similarities between PYO97 and PYO2014*.

The bottom right corner of Figure 4 is populated by PYO97's bins with low bin coverage by PYO2014's reads, whilst the top left clusters PYO2014's bins with low bin coverage by PYO97's reads.

The bins in these two parts of the figure are thus most likely phages added (top left corner) or removed (lower right corner) when constructing the cocktails at the two time points: 1997 and 2014.

We next determined the distances in composition between samples PYO97, PYO2000 and PYO2014 using Mash. The algorithm searches shared k-mers between samples and gave a measure of global mutation distance that takes continuous values between 0 and 1. For each representative k-mer, Mash does not take into account how many of those k-mers are present in each sample, only whether it is present or not. Therefore, the distances are to be considered qualitatively as distances in the variety of phages between samples and not as differences in phage abundances.

Distances are, in general, low between the samples ($D < 0.2$), Table 10, as expected since the different samples are of the same cocktail and contain mostly shared sequences. PYO2014 has the highest distance to the other two samples. From this, it can be derived that a higher number of phages are unique to PYO2014 and absent in the other samples. Conceivably from the date of production, PYO97 and PYO2000 are less distant to each other (0.113 ± 0.0006) than they are to PYO2014, (0.132 ± 0.0008 and 0.138 ± 0.0009, respectively).

Table 10. Global mutation distances between samples.

Sample	PYO97	PYO2000	PYO2014
PYO97	0		
PYO2000	0.113 ± 0.0006	0	
PYO2014	0.132 ± 0.0008	0.138 ± 0.0009	0

4. Discussion

In this paper, we aimed to investigate the composition of four batches of PYO cocktail, produced at the Eliava Institute in 1997, 2000, 2010, and 2014, by means of sequencing and metagenomic analysis. The PYO cocktails from 1997 and 2014 had been stored in a fridge at approximately 4 °C. We were able to extract DNA of high quality from these samples and likewise obtained high-quality sequence reads. We did not test the infectivity of the phages in the cocktails, but have previously found that phages from another cocktail from the Eliava Institute, the INTESTI cocktail, retain their infectivity after storage under similar conditions for at least two years [9]. The phages in the INTESTI cocktail lost their infectivity when they were frozen by mistake without the addition of glycerol. Similarly, the PYO cocktails from 2000 and 2010 had been frozen without the recommended addition of glycerol [31]. Following thawing, we were not able to extract enough DNA of good quality from these cocktails and only obtained sequence reads from PYO2000, which were furthermore of a poorer quality than from PYO97 and PYO2014. We did not test whether the phages in PYO2000 and PYO2010 had also lost their infectivity, but expect that they had. It is worth mentioning that the recommended long-term storage of phages is freezing −80 °C after addition of glycerol [31]. Alternatively, phages can be freeze dried and stored at room temperature [32].

The reads from PYO97 and PYO2014 were assembled into contigs, which were binned into phage draft genomes in a reference independent manner. This is contrary to what was previously done for the INTESTI cocktail [9], where contigs were binned based on Blast searches to public databases. For the purpose of binning the contigs, we used MetaBAT, a method that bins according to the tetranucleotide frequency distances of the contigs. Further, MetaBAT is able to use the co-abundances of contigs in multiple samples, i.e., the consistency in coverage fluctuations of groups of contigs between samples. The method is optimized to handle huge assemblies for a number of samples greater than ten. Since our study involved only two samples of good quality, MetaBAT could not base contigs binning on the co-abundance information, but only on the tetranucleotide frequency distances. This might explain why, consequent to binning, we had to manually curate the generated bins. Two phage draft genomes, one from each sample, were in fact each manually split into two phage draft genomes and other bins were merged according to the coverage consistency and closest reference genome, resulting in five merged draft genomes.

Phage draft genomes were further classified into categories based on their similarity to a reference genome and/or to a phage draft genome in the other sample. This allowed us to identify a group of phage draft genomes that were highly similar to a reference genome and present in both samples. These included draft genomes predicted to target *E. faecalis*, *E. faecium*, *E. coli*, *P. mirabilis*, *P. aeruginosa*, and *S. aureus*. Other near-complete and partial draft genomes, even if without a counterpart in the other sample or reference genome, were predicted to target also *C. sakazakii*, *K. pneumoniae*, *Shigella*, and species of *Salmonella*. Only the prediction of phages targeting *C. sakazakii* and *K. pneumoniae* were counter to our expectations as the declared activity of the PYO cocktail includes *Shigella*, *Salmonella*, *E. coli*, *Proteus*, *S. aureus*, *P. aeruginosa* and *Enterococcus*. Previous studies have shown the close taxonomic relatedness between bacteria of the Enterobacteriaceae family [33,34], which includes *Escherichia*, *Klebsiella*, *Salmonella* and *Shigella*, suggesting that the prediction of *K. pneumoniae* might be a misprediction. Besides, even though phages are usually strain-specific, phages capable of infecting distinct but related hosts, polyvalent phages, are commonly observed among phages of Enterobacteria [35–38], which does not rule out the presence of this type of phages in the cocktail.

To the best of our knowledge, the ANI thresholds for when a phage belongs to a certain species, genus, or family have not been defined. However, we suggest that the phage draft genomes in category 1 and 2 represent phages that likely belong to previously sequenced phage species or at least previously defined genera. Examples include PYO97_11 and PYO2014_26.13 that both closely resemble *Pseudomonas* phage PEV2, a N4likevirus. The phage draft genomes in categories 3 and 4 are, on the other hand, likely to be the first representatives of previously undefined genera, in some cases perhaps even previously undefined sub-families or families, with ANI to the closest reference genomes from 10% to 70%. Examples include PYO97_27.21 which closely resembles PYO2014_28. Both phage draft genomes have an ANI to the closest reference of only approximately 20%. Another example is PYO2014_7, which does not have a counterpart in PYO97 and only has a ANI of 14.3% to the closest reference.

A total of twenty-two new near-complete or partial draft genomes were discovered, which did not resemble any publicly available genomes, or had only poor similarity to one. One of these phage draft genomes was even found to be present in both samples and with high relative abundance (PYO97_27.21/PYO2014_28).

In correspondence to this high number of previously unsequenced phage draft genomes, we also observed a relatively high percentage of reads that could not be mapped to any known phage genome. For PYO97, 11% of the reads could not be mapped to known phage sequences, while the corresponding percentage for PYO2014 was 39%. This relates to the continued scarceness of phage genome representation in public databases compared to bacterial sequences [39]. A previous study from 2013 was able to map 61% of the reads from the Microgen ColiProteus cocktail to public genomes [10].

For PYO97, 17% of the bins were not predicted to be of phage origin, while for PYO2014 the corresponding percentage was 10. When predicting if bins were of phage origin, we used MetaPhinder with a very stringent threshold of 10% ANI. This is a far more conservative threshold than suggested in the original paper describing the MetaPhinder method [23], where the ANI threshold to classify a contig as of phage origin was set to 1.7% ANI. Further, the performance of MetaPhinder is dependent on the size and diversity of a reference database of previously sequenced phages. We thus consider it likely that the bins predicted to be of non-phage origin are due to a limited diversity in the previously sequenced phage genomes rather than, e.g., contamination. This hypothesis is supported by the analysis using MGmapper, which showed that only a negligible amount of the raw sequence reads mapped to reference databases containing sequences from Bacteria, Archaea, MetaHitAssembly, HumanMicrobiome, Bacteria_draft, Human, Virus, or Fungi. Most of the bins predicted to be of non-phage origin had the highest similarity to sequences annotated as uncultured Mediterranean phages. It is worth noticing that phages annotated as uncultured Mediterranean phages counted 28.8%

of the 3889 WGS used to search for references to the bins, which raises the chance that they were randomly selected.

A coverage analysis that included PYO2000 showed a closer similarity of this cocktail batch to the batch from 1997 than that from 2014, in terms of composition. This is also to be expected as there are only 3 years between the production of the first two cocktails compared to the second and the last batches, which were produced with 14 years in between. The phage draft genomes of the PYO97 and PYO2014 cocktails showed huge differences in depth coverages within the samples, indicating as much as a thousand-fold difference between the most and least abundant phages. We speculate that the draft genomes represented by few sequence reads may derive from phages of older batches that have been diluted over time. Alternatively, they may derive from activated prophages integrated in the bacterial hosts used for phage enrichment, as previously suggested [10]. In the previous study by our group of the INTESTI cocktail [9], we did not observe such high differences in abundances. This might, however, be due to the general much lower sequencing depth of the INTESTI cocktail, which would not have allowed for the detection of the phages found at very low concentrations. It is worth pointing out the composition comparison presented here could not account for potential compositional variations within the batches nor for any biases that might have been introduced during sample processing. This is an insight that could be gained by analyzing multiple samples per batch and/or introducing replicates; however, this was beyond the scope of this study.

One of the limitations of the analysis applied here is that neither the lab sample preparation nor the sequencing library construction enriched for RNA sequences. Therefore, likely present *Pseudomonas* phages of potential clinical importance as antimicrobials [40], could not be detected. Besides small RNA coliphages, ssDNA phages were likely missed. In fact, the amplification step of the Illumina sequencing used here is based on the ligation of dsDNA adapters to sheared DNA. Since the ligation occurs between dsDNA fragments, ssDNA phages of the families *Microviridae* and *Inoviridae* could not be efficiently recovered by this approach [41,42]. Furthermore, the binning method that we chose yielded only bins of 10,000 bp or larger. Although we were able to bin more than 90% of the basepairs represented in the contigs, the threshold of 10,000 bp might have sorted out small DNA phages, for instance small *E.coli* phages [43].

5. Conclusions

In the present study, we have performed metagenomic sequencing and analysis of phage cocktails produced over 18 years. Some of the observed phages are common to the phage cocktails and are likely to belong to previously defined phage species and genera. However, we also discovered new phages that only poorly resemble any of the whole genome phage sequences found in public databases. They are likely to represent new genera or even new phage families. For a fuller characterization of the content of the cocktails, methods that also allow for RNA isolation and enrichment and binning processes that allow for the formation of smaller bins, is needed. The raw reads from this study are publicly available at http://www.ebi.ac.uk/ena/data/view/PRJEB23244. The draft genomes have been deposited on the European Nucleotide Archive with accession numbers from ERS1989512 to ERS1989570. It is the authors' hope that this will allow other researchers to continue analyzing and characterizing these phages. The characterization of the cocktail is a first step towards recognizing the PYO cocktail as a regulated drug in western countries.

Supplementary Materials: The following are available online at www.mdpi.com/1999-4915/9/11/328/s1, Table S1: Percentages of PYO97 and PYO2014 reads mapping to MGmapper databases; Table S2: PYO97–Near-complete draft genome with high resemblance to reference phage, but no counterpart in PYO2014; Table S3: PYO2014–Near-complete draft genome with high resemblance to reference phage, but no counterpart in PYO97.

Acknowledgments: We are grateful to Elizabeth Kutter (Olympia, Washington), Zemphira Alavidze and Marina Goderdzishvili (Tbilisi, Georgia) for the PYO cocktails. Thanks to Marlene Dalgaard for providing excellent technical assistance.

Author Contributions: Julia Villarroel extracted the DNA and analyzed the metagenomic data. Mogens Kilstrup contributed reagents and instruments. Julia Villarroel, Mette Voldby Larsen and Morten Nielsen wrote the paper. All authors contributed in reviewing the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

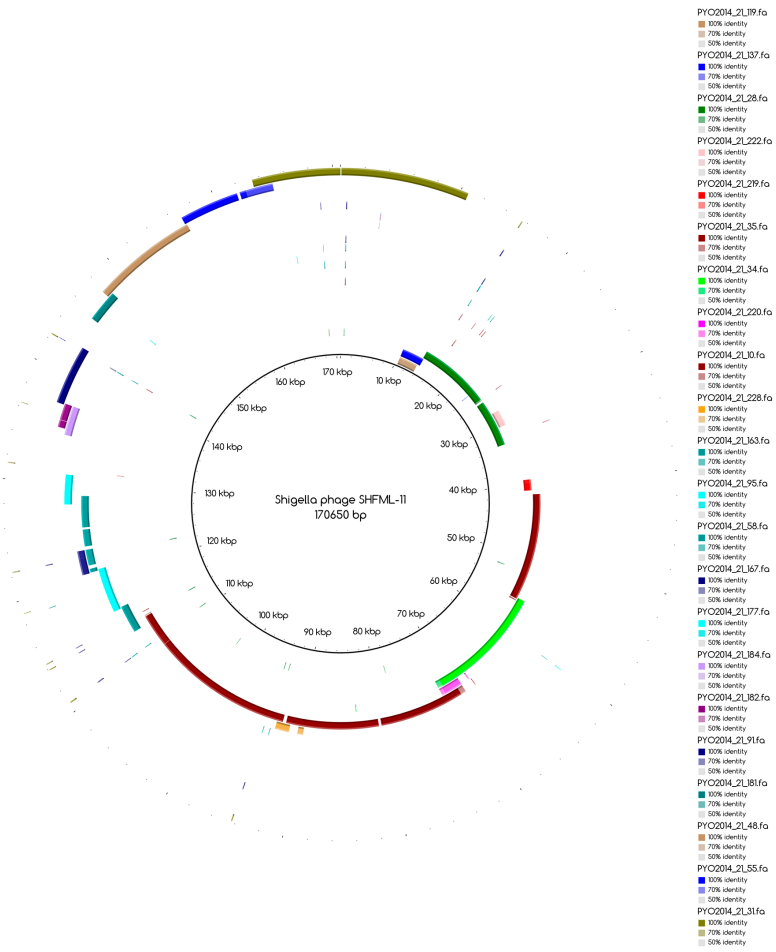


Figure A1. Blast Atlas of PYO2014_21 towards the reference *Shigella* phage SHFML-11.

References

1. McGann, P.; Snesrud, E.; Maybank, R.; Corey, B.; Ong, A.C.; Clifford, R.; Hinkle, M.; Whitman, T.; Lesho, E.; Schaecher, K.E. Escherichia coli Harboring MCR-1 and *bla* $_{\text{CTX-M}}$ on a Novel IncF Plasmid: First report of MCR-1 in the United States. *Antimicrob. Agents Chemother.* **2016**, *60*, 4420–4421. [CrossRef] [PubMed]

2. Liu, Y.-Y.; Wang, Y.; Walsh, T.R.; Yi, L.-X.; Zhang, R.; Spencer, J.; Doi, Y.; Tian, G.; Dong, B.; Huang, X.; et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: A microbiological and molecular biological study. *Lancet Infect. Dis.* **2016**, *16*, 161–168. [CrossRef]
3. WHO | Antimicrobial Resistance. Available online: http://www.who.int/mediacentre/factsheets/fs194/en/ (accessed on 31 August 2017).
4. Twort, F.W. Further investigations on the nature of Ultra-Microscopic viruses and their cultivation. *J. Hyg.* **1936**, *36*, 204–235. [CrossRef] [PubMed]
5. Kutter, E.; de Vos, D.; Gvasalia, G.; Alavidze, Z.; Gogokhia, L.; Kuhl, S.; Abedon, S.T. Phage therapy in clinical practice: Treatment of human infections. *Curr. Pharm. Biotechnol.* **2010**, *11*, 69–86. [CrossRef] [PubMed]
6. Abedon, S.T.; Kuhl, S.J.; Blasdel, B.G.; Kutter, E.M. Phage treatment of human infections. *Bacteriophage* **2011**, *1*, 66–85. [CrossRef] [PubMed]
7. Harper, D.R.; Anderson, J.; Enright, M.C. Phage therapy: Delivering on the promise. *Ther. Deliv.* **2011**, *2*, 935–947. [CrossRef] [PubMed]
8. Vieu, J.F. Intérêt des bactériophages dans le traitement de staphylococcies. *Vie Med.* **1961**, *42*, 823–829. [PubMed]
9. Zschach, H.; Joensen, K.G.; Lindhard, B.; Lund, O.; Goderdzishvili, M.; Chkonia, I.; Jgenti, G.; Kvatadze, N.; Alavidze, Z.; Kutter, E.M.; et al. What can we learn from a metagenomic analysis of a georgian bacteriophage cocktail? *Viruses* **2015**, *7*, 6570–6589. [CrossRef] [PubMed]
10. McCallin, S.; Alam Sarker, S.; Barretto, C.; Sultana, S.; Berger, B.; Huq, S.; Krause, L.; Bibiloni, R.; Schmitt, B.; Reuteler, G.; et al. Safety analysis of a Russian phage cocktail: From metagenomic analysis to oral application in healthy human subjects. *Virology* **2013**, *443*, 187–196. [CrossRef] [PubMed]
11. Babraham Bioinformatics—FastQC, a Quality Control Tool for High Throughput Sequence Data. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 31 August 2017).
12. Schmieder, R.; Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinform. Oxf. Engl.* **2011**, *27*, 863–864. [CrossRef] [PubMed]
13. NIHR Biomedical Research Centre for Mental Health: Computational Biology. Available online: http://compbio.brc.iop.kcl.ac.uk/software/cmpfastq.php (accessed on 31 August 2017).
14. Petersen, T.N.; Lukjancenko, O.; Thomsen, M.C.F.; Maddalena Sperotto, M.; Lund, O.; Møller Aarestrup, F.; Sicheritz-Pontén, T. MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads. *PLoS ONE* **2017**, *12*, e0176469. [CrossRef]
15. Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinform. Oxf. Engl.* **2010**, *26*, 589–595. [CrossRef] [PubMed]
16. Wood, D.E.; Salzberg, S.L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **2014**, *15*, R46. [CrossRef] [PubMed]
17. Nurk, S.; Meleshko, D.; Korobeynikov, A.; Pevzner, P.A. metaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **2017**, *27*, 824–834. [CrossRef] [PubMed]
18. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **2012**, *19*, 455–477. [CrossRef] [PubMed]
19. Kang, D.D.; Froula, J.; Egan, R.; Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **2015**, *3*, e1165. [CrossRef] [PubMed]
20. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
21. NCBI Viral RefSeq Database. Available online: ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/viral/ (accessed on 31 August 2017).
22. PhAnToMe Current Genbank Genomes Downloads. Available online: http://www.phantome.org/Downloads/genomes/genbank/current/ (accessed on 31 August 2017).
23. Jurtz, V.I.; Villarroel, J.; Lund, O.; Voldby Larsen, M.; Nielsen, M. MetaPhinder-identifying bacteriophage sequences in metagenomic data sets. *PLoS ONE* **2016**, *11*, e0163111. [CrossRef] [PubMed]
24. Villarroel, J.; Kleinheinz, K.A.; Jurtz, V.I.; Zschach, H.; Lund, O.; Nielsen, M.; Larsen, M.V. HostPhinder: A phage host prediction tool. *Viruses* **2016**, *8*, 116. [CrossRef] [PubMed]
25. Lee, I.; Kim, Y.O.; Park, S.-C.; Chun, J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evolut. Microbiol.* **2015**. [CrossRef] [PubMed]

26. Alikhan, N.-F.; Petty, N.K.; Ben Zakour, N.L.; Beatson, S.A. BLAST Ring Image Generator (BRIG): Simple prokaryote genome comparisons. *BMC Genom.* **2011**, *12*, 402. [CrossRef] [PubMed]
27. Stothard, P.; Wishart, D.S. Circular genome visualization and exploration using CGView. *Bioinform. Oxf. Engl.* **2005**, *21*, 537–539. [CrossRef] [PubMed]
28. Ondov, B.D.; Treangen, T.J.; Melsted, P.; Mallonee, A.B.; Bergman, N.H.; Koren, S.; Phillippy, A.M. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **2016**, *17*, 132. [CrossRef] [PubMed]
29. Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K.S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; Yamada, T.; et al. A human gut microbial gene catalog established by metagenomic sequencing. *Nature* **2010**, *464*, 59–65. [CrossRef] [PubMed]
30. NIH Human Microbiome Project—Home. Available online: https://www.hmpdacc.org/ (accessed on 31 August 2017).
31. Fortier, L.-C.; Moineau, S. Phage production and maintenance of stocks, including expected stock lifetimes. *Methods Mol. Biol. Clifton N. J.* **2009**, *501*, 203–219. [CrossRef]
32. Blair, J.E.; Williams, R.E.O. Phage typing of staphylococci. *Bull. World Health Organ.* **1961**, *24*, 771–784. [PubMed]
33. Delmas, J.; Breysse, F.; Devulder, G.; Flandrois, J.-P.; Chomarat, M. Rapid identification of Enterobacteriaceae by sequencing DNA gyrase subunit B encoding gene. *Diagn. Microbiol. Infect. Dis.* **2006**, *55*, 263–268. [CrossRef] [PubMed]
34. Hong Nhung, P.; Ohkusu, K.; Mishima, N.; Noda, M.; Monir Shah, M.; Sun, X.; Hayashi, M.; Ezaki, T. Phylogeny and species identification of the family Enterobacteriaceae based on dnaJ sequences. *Diagn. Microbiol. Infect. Dis.* **2007**, *58*, 153–161. [CrossRef] [PubMed]
35. Parra, B.; Robeson, J. Selection of polyvalent bacteriophages infecting Salmonella enterica serovar Choleraesuis. *Electron. J. Biotechnol.* **2016**, *21*, 72–76. [CrossRef]
36. Park, M.; Lee, J.-H.; Shin, H.; Kim, M.; Choi, J.; Kang, D.-H.; Heu, S.; Ryu, S. Characterization and comparative genomic analysis of a novel bacteriophage, SFP10, simultaneously inhibiting both Salmonella enterica and Escherichia coli O157:H7. *Appl. Environ. Microbiol.* **2012**, *78*, 58–69. [CrossRef] [PubMed]
37. Malki, K.; Kula, A.; Bruder, K.; Sible, E.; Hatzopoulos, T.; Steidel, S.; Watkins, S.C.; Putonti, C. Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. *Virol. J.* **2015**, *12*. [CrossRef] [PubMed]
38. Hamdi, S.; Rousseau, G.M.; Labrie, S.J.; Tremblay, D.M.; Kourda, R.S.; Ben Slama, K.; Moineau, S. Characterization of two polyvalent phages infecting Enterobacteriaceae. *Sci. Rep.* **2017**, *7*, 40349. [CrossRef] [PubMed]
39. Dutilh, B.E.; Cassman, N.; McNair, K.; Sanchez, S.E.; Silva, G.G.Z.; Boling, L.; Barr, J.J.; Speth, D.R.; Seguritan, V.; Aziz, R.K.; et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **2014**, *5*, 4498. [CrossRef] [PubMed]
40. Yang, Y.; Lu, S.; Shen, W.; Zhao, X.; Shen, M.; Tan, Y.; Li, G.; Li, M.; Wang, J.; Hu, F.; et al. Characterization of the first double-stranded RNA bacteriophage infecting Pseudomonas aeruginosa. *Sci. Rep.* **2016**, *6*, 38795. [CrossRef] [PubMed]
41. Solonenko, S.A.; Ignacio-Espinoza, J.C.; Alberti, A.; Cruaud, C.; Hallam, S.; Konstantinidis, K.; Tyson, G.; Wincker, P.; Sullivan, M.B. Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genom.* **2013**, *14*, 320. [CrossRef] [PubMed]
42. Székely, A.J.; Breitbart, M. Single-stranded DNA phages: From early molecular biology tools to recent revolutions in environmental microbiology. *FEMS Microbiol. Lett.* **2016**, *363*. [CrossRef] [PubMed]
43. Friedman, S.D.; Genthner, F.J.; Gentry, J.; Sobsey, M.D.; Vinjé, J. Gene mapping and phylogenetic analysis of the complete genome from 30 single-stranded RNA male-specific coliphages (family Leviviridae). *J. Virol.* **2009**, *83*, 11233–11243. [CrossRef] [PubMed]

CHAPTER 4

Characterization of historical phages

Phage typing is a technique in epidemiology, used to identify the strain of pathogenic bacteria. It consists in making a loan of bacteria on a petri dish and dropping a set of well determined phages. After some hours, usually after one night, the sensitivity pattern is observed. Based on which phages are capable of producing plaques, the bacterial strain is determined.

The initial aim of this project was to analyse the phylogenetic relationships between the phages and the respective propagating strains (PSs). The DNA from the PSs has been extracted by Henrike Zschach, then sequenced and phylogenetically analysed in a bachelor project by Saher Munir Shah. Data on the PSs are not presented in this thesis.

Another application that was intended for the results of this study was to improve HostPhinder's host prediction from species to strain level. The susceptibility range of S.aureus strains to the different phages could in fact provide understanding on genomic signatures that determine phage susceptibility at the strain level. An insight that could help improving the resolution of HostPhinder to the strain level, at least for S. aureus phages.

Here I present a manuscript that is till in its early stages. The analyses performed and the preliminary results are described in details with the hope to offer a starting point for further investigations on the dataset.

4.1 Characterisation historical Staphylococcus aureus phages used for phage typing.

Julia Villarroel, Henrike Zschach, Saher Munir Shah, Mogens Kilstrup, Mette Voldby Larsen, Morten Nielsen

4.1.1 introducion

Phage typing, the use of bacteriophages, phages, to identify bacterial strains based on the sensibility pattern of the bacterial sample to a standardized set of phages, is used in epidemiology to identify the source of a disease outbreak. The technique has been widely used to identify pathogenic strains of Staphylococcus aureus, which is a leading cause of nosocomial infections. The spreading of S. aureus multiple-drug-resistant strains is, in fact, of great concern to hospitals worldwide [115, 8]. Since 1953 the International Committee on Systematic Bacteriology strove to coordinate worldwide laboratories that performed phage typing, by providing standardized guidelines on the procedure. In its last official meeting in Brno in 1974, the International Committee proclaimed the Basic International Set of 23 phages for routine typing [10].

```
group I          29, 52, 52A, 79, 80
group II         3A, 3C, 55, 71
group III        6, 42E, 47, 53, 54, 75, 77, 83A, 84, 85
not allocated    81, 94, 95, 96.
```

allowing for the use of additional phages such as 42B, 47C, 52B, 69, 73, and 78 found to be useful locally. Every 4 years, starting from the coordinating efforts in 1953, the International Centre for Staphylococcal Phage-Typing, Colindale, London, distributed newly made freeze-dried stocks of the International Set together with the respective propagating strains (PS). The PS were identified with the same number used to designate the phages; therefore phage 29 was propagated on PS 29, while PS 52A/79 was used to propagate both phages 52A and 79 [10].

The technique, although relatively fast, cheap and allowing to analyse multiple samples at a time, had multiple drawbacks. One of them being the decline in percentage of S. aureus strains susceptible to lyses by the phages in the International Set due to increase in resistance over the years [10]. Besides, the procedures of international coordination and standardization of results interpretation made it a cumbersome technique to use and maintain [71]. Lately, the introduction of DNA-based approaches, such as multilocus sequence typing (MLST), pulsed-field gel electrophoresis (PFGE), and randomly amplified polymorphic DNA analysis (RAPD), ensured greater discriminatory power, better reproducibility and required less labor [48, 13]. As a consequence, molecular typing methods, which allow for real-time monitoring and identification of infectious disease outbreak, are replacing phenotypic characterisation in epidemiology [58, 68, 60, 52, 88, 123]. An exception to this trend though is represented by global surveillance of Salmonella food-borne outbreaks, which is still widely based on phage typing [11]. Eleven of the phages from the basic International Set have been previously sequenced [65], namely 3A, 47, 29, 77, 42E, 55, 52A, 53, 71, 85, and 96.

In this study we sequenced 24 phages which belong to the International Basic Set and their respective propagating strains. We enriched the phages using the respective propagating strain.

4.1.2 Materials and Methods

4.1.2.1 Phage resuspension

Lyophilized phage preparations were resuspended in 1 mL Tryptic Soy Broth (TSB) added 10 mM CaCl2, to facilitate phage infection of the PS and 25 mM MgCl2, for conservation purposes. Ten L of this original stock solution were used for phage titration, while DNA was extracted from the remaining volume.

4.1.2.2 Phage titration and lysate preparation

Ten L of each decinormal serial dilutions from log(0) to log(-6) were dripped on top agar containing 100 L 5h culture of the respective propagating strain. The following day, plaques were observed and 100 L PS and 100 L of the respective phage at the dilution factor which just failed to give confluent plaques were mixed in 4 mL molten top agar and poured on a plate for lysate preparation. The following day the top agar was scraped off with the help of 4 mL TSB and centrifuged to pellet the bacteria. DNA was the extracted from the supernatant, the lysate.

4.1.2.3 Phage DNA extraction and library preparation

The original stock solutions and the lysates were filtered through 22 m filters to remove any bacterial debris. The DNA was extracted and isolated using the NORGEN Phage DNA Isolation Kit (Cat. 46800) following the manual. The extracted DNA was kept at -20oC until library preparation. The DNA was shred into fragments of an average length of 400bp and 250bp and sequenced by Illumina MiSeq generating paired-end reads.

4.1.2.4 Phage Reads Trimming and Assembly

Reads were checked for quality with fastQC [9] and trimmed using Prinseq-lite 0.20.4 [**schmieder_prinseq**] with the following settings: `-trim_qual_right 20 -min_qual_mean 20 -min_len 35 -trim_left 20 -trim_right 20 -derep 123`. Non parallel reads, resulting from trimming, were compensated using cmpfastq [106]. Trimmed reads were assembled using IDBA-UD [83].

4.1.2.5 Best reference genome

The reference genome was searched using MetaPhinder [59]. To this purpose, phage whole genome sequences (WGS) were downloaded from NCBI viral RefSeq database [77] and PhAnToMe [86], resulting in 3,889 unique WGS as of May 2017. MetaPhinder is a Blast[7]-based method designed to discriminate between contigs of phage origin or not. The method has a certain flexibility that allows for arbitrary database and query as far as they are given as fasta files. MetaPhinder calculates the Average Nucleotide Identity (ANI) between a query and a Blast database as follows

$$\mathrm{ANI} = \frac{\sum_{i_1}^{n} \mathrm{id}_i \mathrm{al}_i}{\sum_{i=1}^{n} \mathrm{al}_i} m_{\mathrm{cov}} \tag{4.1}$$

where n is the number of Blastn hits between the query sequence and all sequences in the database, id is the Blastn % identity value between the query and a given database hit, al is the corresponding Blastn alignment length, and m_{cov} is the coverage of the query sequence over all hits. Here a Blast database was constructed from a fasta file containing assembled contigs of the 24 phages. Then the Blast database was searched with the 3,889 phage WGS and the ANI of each WGS was calculated by MetaPhinder. We finally selected the phage genome with the highest ANI as reference genome.

4.1.2.6 Blast alignment visualization

The 24 phage genome were aligned to reference genomes using Blastn. From Blastn results we produced xml files using a customized python script and visualised the circular genomes using the CGView Java Package [107].

4.1.2.7 Prediction of resistance and virulence genes

The presence of resistance was investigated using ResFinder [125]. ResFinder performs a Blastn search of the query sequence against selected databases of publicly available resistant genes. In this study, all available updated databases (as of September 2017) were selected, namely databases of Aminoglycosides, Beta-lactamases, Fluoroquinolone, Fosfomycin, Fusidic Acid, Glycopeptides,
Macrolide-lincosamide-Streptogramin B, Phenicol, Rifampicin, Sulphonamides, and Tetracycline and Trimethoprim. Pair end reads were given as input, the ID threshold was set to 90% and the minimum alignment length to 60%. Virulence genes were searched with VirulenceFinder [58]. Similarly to ResFinder, VirulenceFinder performs a Blastn search against a selected database of virulence genes specific for a selected bacterial species. The database here selected contained updated sequences (as of February 2016) of S. aureus's virulence genes: hlb, hlgABC, tst, lukED, lukFS-PV, etAB, edinABC, aur, splABE, scn, sak, ACME, and enterotoxins A-E, G-O, R, U, Q. The quest was set with 90% ID threshold and 60% minimum alignment length giving pair end reads as input.

4.1.2.8 Host Prediction

We predicted the species of the bacterial host of the 24 phages and 16 lysates using HostPhinder [116].

4.1.3 Results

We checked whether they were still able to infect the respective propagating strains by dropping serial dilutions to a loan of the respective propagating strain. We obtained lysates for 16 of the samples

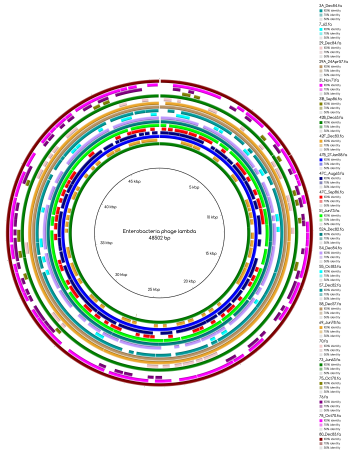
Twenty four samples directly resuspended from the freeze-dried samples together with the 16 lysates were Illumina sequenced. Phages were assembled in contigs and contained between 2 and 827 contigs. For the assembly statistics of each sample, refere to this link : `https://docs.google.com/spreadsheets/d/1tGP0RU5zdcSIxX_evm4EBN1eXESFg-zUd07pAp0Tf9c/edit?usp=sharing`

4.1.3.1 Prediction of resistance and virulence genes

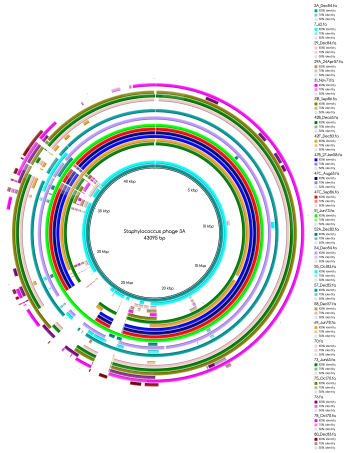
We predicted resistence and virulence genes using the respective webservers at `https://cge.cbs.dtu.dk/services/`. Partial sequence of the gene blaTEM-116, responsible for beta-lactam resistance was found in five samples: 29, 29A, 31B, 42F, and 47C. The location of the gene either at the beginning or at the end of the contigs explains why the sequence of the gene found was not complete. The virulence genes found encode for Beta-homolysin, Leukocidin D component, Panton Valentine leukocidin F component, Staphylokinase, Staphylococcal complement inhibitor, Serine proteases, Toxin shock syndrome toxin-1, and Enterotoxin G,M, O, and U.

4.1.3.2 Best reference genome

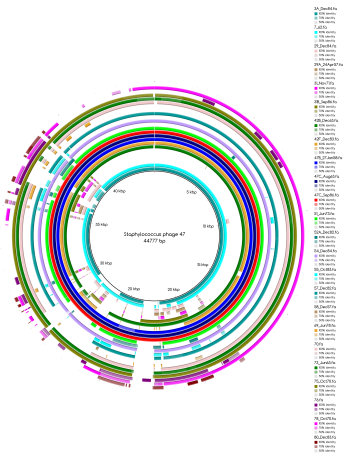
We ran MetaPhinder to find the best reference for the twenty four phages. Curiously, the best reference genome was predicted to be Enterobacteria phage lambda (NC_001416.1) with 99.921% ANI, followed by *Staphylococcus* phage 3A (NC_007053.1), 96.296% ANI, and *Staphylococcus* phage 47 (NC_007054.1), 96.27% ANI. See 4.1.



(a) Enterobacteria phage lambda



(b) *Staphylococcus* phage 3A



(c) *Staphylococcus* phage 47

Figure 4.1: Blast Atlases showing the alignment of the 24 phages to the reference genome (**a**) Enterobacteria phage lambda, (**b**) *Staphylococcus* phage 3A and (**c**) *Staphylococcus* phage 47 .

4.1.3.3 Host Prediction

Almost all phages and lysates were predicted to infect *S. aureus* with a coverage value > 0.7, which is considered to give highly reliable results [116]. Φ31B, Φ69, Φ70, and Φ78 were also predicted to infect *S. aureus*, but with a lower, still reliable, coverage. The Φ31B obtained from the lysate has also a lower coverage while no lysate were

obtainable from the other 3 phages. Φ58 was predicted to infect *Shigella flexneri* with a coverage of 0.21, which is still considered to give reliable results. We were not able to obtain plaques of this phage on the respective PS.

CHAPTER 5

Conclusion and future remarks

Due to the rapid spread of antibiotic resistance all around the world, phages have become subject of great interest as alternative or supplementary therapy to antibiotics. Institutes in the former Eastern Blocs, such as the Georgian Eliava Institute in Tbilisi, have almost a hundred years of experience in treating patients with phages. The combination of this practical knowledge with rigorous scientific methods for testing and proving the efficacy of phage therapy through double-blind, placebo-controlled clinical trials, could help introduce these natural bacterial killers in the western clinical practice. The work presented in this dissertation is divisible in two main topics: the prediction of the bacterial host, presented in HostPhinder, and a second part focused on NGS tools for characterizing phage metagenomes and single genomes. This included the sequencing of PYO phage cocktails and of historical phages used for phage typing.

In the first project, the aim was to predict the bacterial host of phages based on their entire genome sequence. The tool dubbed HostPhinder was able to predict the host species of an independent evaluation set with more than 74% accuracy. Since its publication in May 2016, the paper describing the HostPhinder method has been cited 5 times and the WebServer, available at `https://cge.cbs.dtu.dk/services/HostPhinder/`, has been ran 885 times, as of 16th October 2017. As an example, HostPhinder has been used for the prediction of the host of a phage-like plasmid [27] and has been suggested as a complementary tool to predict the bacterial host of predicted phage sequences in metagenomes [81, 33]. Thanks to widely available High Throughput Sequencing (HTS) methods, metagenomics has become the most effective and comprehensive approach for the genomic analysis of uncultured microbial populations [37]. The limiting step is extracting useful information about community function, phylogeny, evolution, and associations between biological entities. More than 8,000 metagenome datasets are nowaday available and yet not completely characterized [54]. In the realm of phage discovery, a next step could be developing a pipeline for phage identification and characterization. This would give insight into the phylogeny and evolutionary profiles as well the discovery of phage-host linkages. The first step would be the detection of viral sequences. This task can be performed based on similarity to known viral sequence. An example of such reference-based methods is

given by MetaPhinder [59], which calculates an average nucleotide identity to publicly available sequences based on hits obtained using Blast. Another interesting approach for finding viral sequences is given by data-driven association studies between functional entities in metagenomes. In one of such studies, Nielsen et al. [79] clustered genes based on trends of co-abundance across multiple metagenomics samples. They further discovered associations between small co abundance gene groups (CAGs) and big ones, metagenomic species (MGSs). This association was such that small CAGs were only present in the presence of certain MGSs. The authors suggested that the small CAGs could represent phages. The detection of viral sequences together with the prediction of the host can aid in the identification of potential phage therapy treatments [99, 81]. Furthermore, the finding of associations through the data-driven approach described before, can give us insights upon the functional and evolutionary patterns of bacteria-phage communities. Such understandings can be applied, for instance, in engineering of microbial communities, where phages can be used as tools for the modulation of the gut microbiome. HostPhinder by predicting the bacterial host, represents the first step of characterization of metagenomically-derived viruses. New methods have been developed to predict phage hosts. The VirHostMatcher software [5] bases phage host prediction on oligonucleotide frequencies (ONF) distances between host and viral sequences. The ONF usage may in fact be driven by evolutionary pressure on the virus to avoid sequences recognized by host restriction enzymes [104, 89]. Another phage host prediction tool, WhiSh [43] bases the prediction on the highest likelihood of a query phage sequence estimated under homogeneous Markov models trained on potential host genomes. The future will tell which of these different approaches or their combinations will be most successful in accurately predicting the host of phages.

The second study presents a metagenomic characterization of 3 batches of the Georgian PYO phage cocktail. Using HostPhinder, we found phage representatives of all bacterial targets supposedly covered by the cocktail's. Furthermore, by comparing the composition of the cocktails, we demonstrated that it remained relatively stable throughout the years. Also our results strongly suggest the presence of a novel previously uncharacterized phage present in the different batches of the PYO cocktail. Other studies have aimed at assessing the in vitro spectrum of susceptibility of bacterial strains of human origin to the Georgian cocktail. In particular the PYO cocktail's activity was tested against multiple strains of *E. coli*, *Proteus spp.* and *P. aeruginosa* [49, 16, 39], showing promising results of activity against multidrug resistant (MDR) pathogens. The in vitro efficacy studies together with the NGS analysis here provided could aim at facilitate the acceptance of this cocktail as western pharmaceutical.

In the last project, we have been working on 24 phages used by the Statens Serum Institute (SSI) for phage typing. The DNA of the respective propagating strains (PS) have also been extracted and sequenced as well as phylogenetically analyzed. The phages and the PSs could be further analyzed phylogenetically in other to investigate trends of co-evolution between phage and host. Another application of this study is the investigation of genomic features that determine phage sensibility at the host strain level. This will enhance the taxonomic resolution of HostPhinder from species

to strain prediction. The further characterization of these phage genomes may aid in finding optimal lytic candidates for phage therapy. A task that is far from trivial since most published *S. aureus* phages belong to the *Siphoviridae* family of temperate phages [18, 46]. The presence of unwanted resistance and virulence genes was detected in some of the phages as described in Chapter 4. The next step would be searching the phage genomes for integrase genes, which are linked to the lysogenic cycle of phages. In the cure against pathogenic bacteria, in fact, not temperate but lytic phages are desirable. *S. aureus* phages can be used as prophylactic treatment against prosthetic joint infections (PJI). In Denmark, Mogens Kilstrup and collaborators are testing lytic *S.aureus* phages in murine models of prosthesis implantations. Here the surface of the prosthesis is sprayed with a solution of the phage prior to implantation.

Phages are important for human health and can be used as therapy against bacterial infections. The PYO cocktail here characterized is an example of a potential pharmaceutical for the cure of skin infection. The sequencing of historical *S.aureus* phages may aid the discovery of candidate phage for additional therapeutical purposes. Finally, phages are highly abundant in the human gut microbiome and have be proven to have a modulatory role that has impact on human health and immune system [47]. Detection of viral sequences coupled with the prediction of the host, with tools like HostPhinder, can provide a better understanding of their modulatory role as well as finding potential therapeutic phages.

Bibliography

[1] Takashi Abe et al. “Informatics for unveiling hidden genome signatures”. In: *Genome research* 13.4 (2003), pages 693–702.

[2] Stephen T Abedon et al. “Phage treatment of human infections”. In: *Bacteriophage* 1.2 (2011), pages 66–85.

[3] Hans-W Ackermann. “Phage classification and characterization”. In: *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions* (2009), pages 127–140.

[4] H-W Ackermann. “5500 Phages examined in the electron microscope”. In: *Archives of virology* 152.2 (2007), pages 227–243.

[5] Nathan A Ahlgren et al. “Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences”. In: *Nucleic acids research* 45.1 (2016), pages 39–53.

[6] Johannes Alneberg et al. “Binning metagenomic contigs by coverage and composition”. In: *Nature methods* 11.11 (2014), pages 1144–1146.

[7] Stephen F Altschul et al. “Basic local alignment search tool”. In: *Journal of molecular biology* 215.3 (1990), pages 403–410.

[8] Lerma F Alvarez et al. “Staphylococcus aureus nosocomial infections in critically ill patients admitted in intensive care units”. In: *Medicina clinica* 126.17 (2006), pages 641–646.

[9] Simon Andrews. “FastQC: a quality control tool for high throughput sequence data”. In: (2010). URL: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc`.

[10] Elizabeth H. Asheshov and Phyllis M. Rountree. “Report (1970-1974) of the Subcommittee on Phage-Typing of Staphylococci to the International Committee on Systematic Bacteriology”. In: *International Journal of Systematic Bacteriology* 25 (1975), pages 241–242.

[11] Dorte Lau Baggesen et al. “Phage typing of Salmonella Typhimurium-is it still a useful tool for surveillance and outbreak investigation”. In: *Euro Surveill* 15.4 (2010), page 19471.

[12] Anton Bankevich et al. “SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing”. In: *Journal of computational biology* 19.5 (2012), pages 455–477.

[13] Tammy L Bannerman et al. “Pulsed-field gel electrophoresis as a replacement for bacteriophage typing of Staphylococcus aureus.” In: *Journal of Clinical Microbiology* 33.3 (1995), pages 551–555.

[14] Rodolphe Barrangou et al. “CRISPR provides acquired resistance against viruses in prokaryotes”. In: *Science* 315.5819 (2007), pages 1709–1712.

[15] Cecilia Bebeacua et al. “Structure, adsorption to host, and infection mechanism of virulent lactococcal phage p2”. In: *Journal of virology* 87.22 (2013), pages 12302–12312.

[16] Odette J Bernasconi et al. “In vitro activity of three commercial bacteriophage cocktails against multidrug-resistant Escherichia coli and Proteus spp. strains of human and non-human origin”. In: *Journal of global antimicrobial resistance* 8 (2017), pages 179–185.

[17] Kyle Bibby. “Improved bacteriophage genome data is necessary for integrating viral and bacterial ecology”. In: *Microbial ecology* 67.2 (2014), pages 242–244.

[18] John E Blair and Miriam Carr. “Lysogeny in staphylococci”. In: *Journal of bacteriology* 82.6 (1961), pages 984–993.

[19] David E Bradley. “Ultrastructure of bacteriophage and bacteriocins.” In: *Bacteriological reviews* 31.4 (1967), page 230.

[20] Arthur Brady and Steven L Salzberg. “Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models”. In: *Nature methods* 6.9 (2009), pages 673–676.

[21] Mya Breitbart et al. “Genomic analysis of uncultured marine viral communities”. In: *Proceedings of the National Academy of Sciences* 99.22 (2002), pages 14250–14255.

[22] Harald Brüssow. “What is needed for phage therapy to become a reality in Western medicine?” In: *Virology* 434.2 (2012), pages 138–142.

[23] Josué L Castro-Mejía et al. “Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut”. In: *Microbiome* 3.1 (2015), page 64.

[24] Hamidreza Chitsaz et al. “Efficient de novo assembly of single-cell bacterial genomes from short-read data sets”. In: *Nature biotechnology* 29.10 (2011), pages 915–921.

[25] Marie-Christine Chopin, Alain Chopin, and Elena Bidnenko. “Phage abortive infection in lactococci: variations on a theme”. In: *Current opinion in microbiology* 8.4 (2005), pages 473–479.

[26] Scott C Clark et al. “ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies”. In: *Bioinformatics* 29.4 (2013), pages 435–443.

[27] Anna Colavecchio et al. “AnCo3, a New Member of the Emerging Family of Phage-Like Plasmids”. In: *Genome announcements* 5.19 (2017), e00110–17.

[28] Phillip EC Compeau, Pavel A Pevzner, and Glenn Tesler. “How to apply de Bruijn graphs to genome assembly”. In: *Nature biotechnology* 29.11 (2011), pages 987–991.

[29] *Competing for Talent in a Global Market*. Blog. 2017. URL: `https://blogs.adobe.com/documentcloud/competing-for-talent-in-a-global-market/`.

[30] Stephanie A Connon and Stephen J Giovannoni. “High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates”. In: *Applied and environmental microbiology* 68.8 (2002), pages 3878–3885.

[31] Eamonn P Culligan et al. “Metagenomics and novel gene discovery: promise and potential for novel therapeutics”. In: *Virulence* 5.3 (2014), pages 399–412.

[32] Brigid M Davis et al. “CTX prophages in classical biotype Vibrio cholerae: functional phage genes but dysfunctional phage genomes”. In: *Journal of bacteriology* 182.24 (2000), pages 6992–6998.

[33] Jonathan Deaton, Feiqiao Yu, and Stephen Quake. “PhaMers identifies novel bacteriophage sequences from thermophilic hot springs”. In: *bioRxiv* (2017), page 169672.

[34] Patrick J Deschavanne et al. “Genomic signature: characterization and classification of species assessed by chaos game representation of sequences.” In: *Molecular biology and evolution* 16.10 (1999), pages 1391–1399.

[35] Felix d’Herelle. “Sur un microbe invisible antagoniste des bacilles dysentériques”. In: *CR Acad. Sci. Paris* 165 (1917), pages 373–375.

[36] Gregory J Dick et al. “Community-wide analysis of microbial genome sequence signatures”. In: *Genome biology* 10.8 (2009), R85.

[37] Melissa B Duhaime and Matthew B Sullivan. “Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline”. In: *Virology* 434.2 (2012), pages 181–186.

[38] Bas E Dutilh et al. “A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes”. In: *Nature communications* 5 (2014).

[39] Christiane Essoh et al. “The susceptibility of Pseudomonas aeruginosa strains from cystic fibrosis patients to bacteriophages”. In: *PLoS One* 8.4 (2013), e60575.

[40] Walter Fiers et al. “Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene”. In: *Nature* 260.5551 (1976), pages 500–507.

[41] Laura S Frost et al. “Mobile genetic elements: the agents of open source evolution”. In: *Nature reviews. Microbiology* 3.9 (2005), page 722.

[42] Dottore Emiliano Fruciano and Shawna Bourne. “Phage as an antimicrobial agent: d’Herelle’s heretical theories and their role in the decline of phage prophylaxis in the West”. In: *Canadian Journal of Infectious Diseases and Medical Microbiology* 18.1 (2007), pages 19–26.

[43] Clovis Galiez et al. “WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs”. In: *Bioinformatics* 33.19 (2017), pages 3113–3114.

[44] Andrew J Gentles and Samuel Karlin. “Genome-scale compositional comparisons in eukaryotes”. In: *Genome research* 11.4 (2001), pages 540–546.

[45] Mohammadreza Ghodsi et al. “De novo likelihood-based measures for comparing genome assemblies”. In: *BMC research notes* 6.1 (2013), page 334.

[46] Christiane Goerke et al. “Diversity of prophages in dominant Staphylococcus aureus clonal lineages”. In: *Journal of bacteriology* 191.11 (2009), pages 3462–3468.

[47] Andrzej Górski et al. “2 Phage as a Modulator of Immune Responses: Practical Implications for Phage Therapy”. In: *Advances in virus research* 83 (2012), page 41.

[48] Hajo Grundmann et al. “Determining the genetic structure of the natural population of Staphylococcus aureus: a comparison of multilocus sequence typing with pulsed-field gel electrophoresis, randomly amplified polymorphic DNA analysis, and phage typing”. In: *Journal of clinical microbiology* 40.12 (2002), pages 4544–4546.

[49] Aycan Gundogdu, Darajen Bolkvadze, and Huseyin Kilic. “In vitro Effectiveness of Commercial Bacteriophage Cocktails on Diverse Extended-Spectrum Beta-Lactamase Producing Escherichia coli Strains”. In: *Frontiers in microbiology* 7 (2016).

[50] Ernest Hanbury Hankin. “L’action bactericide des eaux de la Jumna et du Gange sur le vibrion du cholera”. In: *Ann. Inst. Pasteur* 10.5 (1896), page ll.

[51] DR Harper, J Anderson, and MC Enright. “Phage therapy: delivering on the promise”. In: *Therapeutic delivery* 2.7 (2011), pages 935–947.

[52] Simon R Harris et al. “Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study”. In: *The Lancet infectious diseases* 13.2 (2013), pages 130–136.

[53] Henrik Hasman et al. “Rapid whole genome sequencing for the detection and characterization of microorganisms directly from clinical samples”. In: *Journal of clinical microbiology* (2013), JCM–02452.

[54] Stephen Hayes et al. “Metagenomic Approaches to Assess Bacteriophages in Various Environmental Niches”. In: *Viruses* 9.6 (2017), page 127.

[55] Alfred D Hershey and Martha Chase. “Independent functions of viral protein and nucleic acid in growth of bacteriophage”. In: *The Journal of general physiology* 36.1 (1952), pages 39–56.

[56] Uwe Hobohm et al. “Selection of representative protein data sets”. In: *Protein Science* 1.3 (1992), pages 409–417.

[57] Michael Imelfort et al. “GroopM: an automated tool for the recovery of population genomes from related metagenomes”. In: *PeerJ* 2 (2014), e603.

[58] Katrine Grimstrup Joensen et al. “Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli”. In: *Journal of clinical microbiology* 52.5 (2014), pages 1501–1510.

[59] Vanessa Isabell Jurtz et al. “MetaPhinder—Identifying Bacteriophage Sequences in Metagenomic Data Sets”. In: *PloS one* 11.9 (2016), e0163111.

[60] Rolf S Kaas et al. “Solving the problem of comparing whole bacterial genomes across different sequencing platforms”. In: *PLoS One* 9.8 (2014), e104984.

[61] Dongwan D Kang et al. “MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities”. In: *PeerJ* 3 (2015), e1165.

[62] Fredrik Karlsson et al. “The mechanism of bacterial infection by filamentous phages involves molecular interactions between TolA and phage protein 3 domains”. In: *Journal of bacteriology* 185.8 (2003), pages 2628–2634.

[63] Andrew MQ King et al. *Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses*. Elsevier, 2011.

[64] Elizabeth Kutter et al. “Phage therapy in clinical practice: treatment of human infections”. In: *Current pharmaceutical biotechnology* 11.1 (2010), pages 69–86.

[65] Tony Kwan et al. “The complete genomes and proteomes of 27 Staphylococcus aureus bacteriophages”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.14 (2005), pages 5174–5179.

[66] Eric S Lander et al. “Initial sequencing and analysis of the human genome”. In: (2001).

[67] Mette V Larsen et al. “Benchmarking of methods for genomic taxonomy”. In: *Journal of clinical microbiology* 52.5 (2014), pages 1529–1539.

[68] Pimlapas Leekitcharoenphon et al. “Evaluation of whole genome sequencing for outbreak detection of Salmonella enterica”. In: *PloS one* 9.2 (2014), e87991.

[69] Dinghua Li et al. “MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph”. In: *Bioinformatics* 31.10 (2015), pages 1674–1676.

[70] Heng Li and Richard Durbin. “Fast and accurate short read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* 25.14 (2009), pages 1754–1760.

[71] RR Marples and VT Rosdahl. “International quality control of phage typing of Staphylococcus aureus”. In: *Journal of medical microbiology* 46.6 (1997), pages 511–516.

[72] Michael L Metzker. “Sequencing technologies–the next generation”. In: *Nature reviews. Genetics* 11.1 (2010), page 31.

[73] Samuel Minot et al. “Hypervariable loci in the human gut virome”. In: *Proceedings of the National Academy of Sciences* 109.10 (2012), pages 3962–3966.

[74] Samuel Minot et al. “Rapid evolution of the human gut virome”. In: *Proceedings of the National Academy of Sciences* 110.30 (2013), pages 12450–12455.

[75] Samuel Minot et al. “The human gut virome: inter-individual variation and dynamic response to diet”. In: *Genome research* 21.10 (2011), pages 1616–1625.

[76] John L Mokili, Forest Rohwer, and Bas E Dutilh. “Metagenomics and future perspectives in virus discovery”. In: *Current opinion in virology* 2.1 (2012), pages 63–77.

[77] *NCBI NCBI viral RefSeq database*. `ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/viral/`. Accessed: 2017-05-8.

[78] Melody N Neely and David I Friedman. “Arrangement and functional identification of genes in the regulatory region of lambdoid phage H-19B, a carrier of a Shiga-like toxin”. In: *Gene* 223.1 (1998), pages 105–113.

[79] H Bjørn Nielsen et al. “Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes”. In: *Nature biotechnology* 32.8 (2014), pages 822–828.

[80] Sergey Nurk et al. “metaSPAdes: a new versatile metagenomic assembler”. In: *Genome Research* 27.5 (2017), pages 824–834.

[81] David Paez-Espino et al. “Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data”. In: *nature protocols* 12.8 (2017), pages 1673–1682.

[82] Yu Peng et al. “IDBA–a practical iterative de Bruijn graph de novo assembler”. In: *Annual international conference on research in computational molecular biology*. Springer. 2010, pages 426–440.

[83] Yu Peng et al. “IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth”. In: *Bioinformatics* 28.11 (2012), pages 1420–1428.

[84] Thomas Nordahl Petersen et al. “MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads”. In: *PloS one* 12.5 (2017), e0176469.

[85] Paul A Pevzner, Haixu Tang, and Glenn Tesler. “De novo repeat classification and fragment assembly”. In: *Genome research* 14.9 (2004), pages 1786–1796.

[86] *PhAnToMe PhAnToMe current genbank genomes Downloads*. `http://www.phantome.org/Downloads/genomes/genbank/current/`. Accessed: 2017-05-8.

[87] Guy Plunkett et al. “Sequence of Shiga Toxin 2 Phage 933W fromEscherichia coli O157: H7: Shiga Toxin as a Phage Late-Gene Product”. In: *Journal of bacteriology* 181.6 (1999), pages 1767–1778.

[88] Lance B Price et al. “Staphylococcus aureus CC398: host adaptation and emergence of methicillin resistance in livestock”. In: *MBio* 3.1 (2012), e00305–11.

[89] David T Pride et al. “Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses”. In: *BMC genomics* 7.1 (2006), page 8.

[90] Theodore T Puck and Howard H Lee. “Mechanism of cell wall penetration by viruses”. In: *Journal of Experimental Medicine* 101.2 (1955), pages 151–175.

[91] DV Rakhuba et al. “Bacteriophage receptors, mechanisms of phage adsorption and penetration into host cell”. In: *Pol. J. Microbiol* 59.3 (2010), pages 145–155.

[92] Michael S Rappé and Stephen J Giovannoni. “The uncultured microbial majority”. In: *Annual Reviews in Microbiology* 57.1 (2003), pages 369–394.

[93] Alejandro Reyes et al. “Gnotobiotic mouse model of phage–bacterial host dynamics in the human gut”. In: *Proceedings of the National Academy of Sciences* 110.50 (2013), pages 20236–20241.

[94] Alejandro Reyes et al. “Going viral: next generation sequencing applied to human gut phage populations”. In: *Nature Reviews. Microbiology* 10.9 (2012), page 607.

[95] Alejandro Reyes et al. “Viruses in the faecal microbiota of monozygotic twins and their mothers”. In: *Nature* 466.7304 (2010), pages 334–338.

[96] Corinna Richter, James T Chang, and Peter C Fineran. “Function and regulation of clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR associated (Cas) systems”. In: *Viruses* 4.10 (2012), pages 2291–2311.

[97] Forest Rohwer. “Global phage diversity”. In: *Cell* 113.2 (2003), page 141.

[98] Simon Roux et al. “Ecogenomics and biogeochemical impacts of uncultivated globally abundant ocean viruses”. In: *bioRxiv* (2016), page 053090.

[99] George PC Salmond and Peter C Fineran. “A century of the phage: past, present and future”. In: *Nature reviews. Microbiology* 13.12 (2015), page 777.

[100] Rickard Sandberg et al. “Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier”. In: *Genome research* 11.8 (2001), pages 1404–1409.

[101] Frederick Sanger, Steven Nicklen, and Alan R Coulson. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the national academy of sciences* 74.12 (1977), pages 5463–5467.

[102] Frederick Sanger et al. “Nucleotide sequence of bacteriophage φX174 DNA”. In: *nature* 265.5596 (1977), pages 687–695.

[103] Robert Schmieder and Robert Edwards. “Quality control and preprocessing of metagenomic datasets”. In: *Bioinformatics* 27.6 (2011), pages 863–864.

[104] Paul M Sharp, Mark S Rogers, and David J McConnell. “Selection pressures on codon usage in the complete genome of bacteriophage T7”. In: *Journal of molecular evolution* 21.2 (1985), pages 150–160.

[105] David A. Shub. “Bacterial Viruses: Bacterial altruism?” In: *Current Biology* 4.6 (1994), pages 555–556.

[106] David To Stephen Newhouse. “cmpfastq: A simple perl program that allows the user to compare QC filtered fastq files”. In: (2011). URL: `http://compbio.brc.iop.kcl.ac.uk/software/cmpfastq.php`.

[107] Paul Stothard and David S Wishart. “Circular genome visualization and exploration using CGView”. In: *Bioinformatics* 21.4 (2004), pages 537–539.

[108] Klaus Strebhardt and Axel Ullrich. “Paul Ehrlich’s magic bullet concept: 100 years of progress”. In: *Nature Reviews Cancer* 8.6 (2008), pages 473–480.

[109] Marc Strous et al. “The binning of metagenomic contigs for microbial physiology of mixed cultures”. In: *Frontiers in microbiology* 3 (2012).

[110] William C Summers. “Bacteriophage therapy”. In: *Annual Reviews in Microbiology* 55.1 (2001), pages 437–451.

[111] Hanno Teeling et al. “Application of tetranucleotide frequencies for the assignment of genomic fragments”. In: *Environmental microbiology* 6.9 (2004), pages 938–947.

[112] Hanno Teeling et al. “TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences”. In: *BMC bioinformatics* 5.1 (2004), page 163.

[113] Torsten Thomas, Jack Gilbert, and Folker Meyer. “Metagenomics-a guide from sampling to data analysis”. In: *Microbial informatics and experimentation* 2.1 (2012), page 3.

[114] Frederick W Twort. “An investigation on the nature of ultra-microscopic viruses.” In: *The Lancet* 186.4814 (1915), pages 1241–1243.

[115] Rea Valaperta et al. “Staphylococcus aureus nosocomial infections: the role of a rapid and low-cost characterization for the establishment of a surveillance system”. In: *New Microbiologica* 33.3 (2010), pages 223–232.

[116] Julia Villarroel et al. “HostPhinder: a phage host prediction tool”. In: *Viruses* 8.5 (2016), page 116.

[117] Matthew K Waldor and John J Mekalanos. “Lysogenic conversion by a filamentous phage encoding cholera toxin”. In: *Science* 272.5270 (1996), page 1910.

[118] Steven W Wilhelm and Curtis A Suttle. “Viruses and nutrient cycles in the sea: viruses play critical roles in the structure and function of aquatic food webs”. In: *Bioscience* 49.10 (1999), pages 781–788.

[119] K Eric Wommack et al. “VIROME: a standard operating procedure for analysis of viral metagenome sequences”. In: *Standards in genomic sciences* 6.3 (2012), page 421.

[120] Derrick E Wood and Steven L Salzberg. “Kraken: ultrafast metagenomic sequence classification using exact alignments”. In: *Genome biology* 15.3 (2014), R46.

[121] *World Health Organization Antimicrobial resistance*. `http://www.who.int/mediacentre/factsheets/fs194/en/`. Accessed: 2017-07-28.

[122] Yu-Wei Wu et al. “MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm”. In: *Microbiome* 2.1 (2014), page 26.

[123] Bernadette C Young et al. “Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease”. In: *Proceedings of the National Academy of Sciences* 109.12 (2012), pages 4550–4555.

[124] RY Young. “Bacteriophage lysis: mechanism and regulation.” In: *Microbiological reviews* 56.3 (1992), pages 430–481.

[125] Ea Zankari et al. “Identification of acquired antimicrobial resistance genes”. In: *Journal of antimicrobial chemotherapy* 67.11 (2012), pages 2640–2644.

[126] Daniel R Zerbino and Ewan Birney. “Velvet: algorithms for de novo short read assembly using de Bruijn graphs”. In: *Genome research* 18.5 (2008), pages 821–829.

[127] Henrike Zschach et al. “What can we learn from a metagenomic analysis of a Georgian bacteriophage cocktail?” In: *Viruses* 7.12 (2015), pages 6570–6589.

{2.7182818284}ϑφερτυθιοπσδφγηξκλ